

UNIVERSITÀ DEGLI STUDI DI NAPOLI

“FEDERICO II”

Scuola di Dottorato in Medicina Molecolare

Dottorato di Ricerca in Genetica e Medicina Molecolare



**“TRANSCRIPTOME DISCOVERY IN EMBRYONIC STEM
CELLS: A POST-GENOMIC APPROACH”**

**Coordinatore:
Prof. Carmelo Bruno Bruni**

**Candidato:
Dott. ssa Pamela Claudiani**

Anno

2007

UNIVERSITÀ DEGLI STUDI DI NAPOLI
“FEDERICO II”

“Telethon Institute of Genetics and Medicine (TIGEM)”

Dottorato di Ricerca in Genetica e Medicina Molecolare

Coordinatore Prof. Carmelo Bruno Bruni

Sede amministrativa:
Dipartimento di Biologia e Patologia Cellulare e Molecolare
“Luigi Califano”

UNIVERSITÀ DEGLI STUDI DI NAPOLI

“FEDERICO II”

“Telethon Institute of Genetics and Medicine (TIGEM)”

**Tesi di Dottorato di Ricerca in Genetica e Medicina Molecolare
XIX ciclo**

**“TRANSCRIPTOME DISCOVERY IN EMBRYONIC STEM
CELLS: A POST-GENOMIC APPROACH”**

Candidato: Pamela Claudiani

Docente guida: Elia Stupka

To my family, to Manuel and particularly to my parents:

I know they are proud of me.

Acknowledgements

No words are enough to thank my supervisor Elia Stupka for giving me the opportunity to work in his laboratory, for his assistance, interest, support and encouragement and finally for the preparation of my thesis. I'm very thankful to him for his trust and for letting me free to follow my ideas in every moment.

I gratefully acknowledge Dr. Guglielmo Roma and Dr. Remo Sanges for friendship and for helping me in all my difficulties.

I would like to thanks Gilda Cobellis for providing us some reagents, for being always very kind to me, for her esteem, suggestions and helpful discussions.

Many thanks to all the people I met at the TIGEM institute, but particularly to Giampiero, Vincenza, Marco D, Marco S, Santos, Vincenzo Alessandro, Maria, Francesco, for friendship, support and many interesting discussions (scientific and non-scientific) and for making my permanence at TIGEM very pleasant and fruitful. I'm very grateful to all the researchers at TIGEM for their esteem and for supporting me during my permanence at TIGEM. In particular I acknowledge Dr. ssa Caterina Missero for being always very kind to me, for her esteem and her support.

A special thank-you to Giampiero for his support especially in the last period, for encouraging me and helping me in all technical troubles. Additional thanks to Francesco and Vincenzo Alessandro for listening and having something comforting to say when I was depressed.

I acknowledge Eva Kallmar and Prof. Ferenc Muller from Germany for teaching me the co-injection method in zebrafish and for providing us some reagents.

No thanks can be enough to all the people in the lab, especially to Danilo, Marco, Carmen, Ivana, Christian, Marca for their help and their support.

Finally, I'm deeply grateful to my family for their love and their constant support throughout this period.

Abstract

Since the publication of the human and mouse genomes, several efforts have been undertaken to elucidate not only their coding gene content, but also the full catalogue of other functional non-coding elements contained within them. During my PhD I contributed to the characterization of a novel set of 8,000 genes, prevalently non-coding, and a novel set of 20,000 enhancer elements.

In order to identify novel genes within the mouse genome we used the gene trapping approach in ES cells. Embryonic stem (ES) cells are pluripotent cells with the capacity of self-renewal and the ability to differentiate into specific cell lineages. In this work was performed the first genome-wide analysis of the mouse ES cell transcriptome using 250,000 gene trap sequence tags deposited in all available public databases. We identified >8,000 novel transcripts of which a great part revealed as non-coding, and >1,000 novel alternative and often tissues specific exons of known genes. We validate experimentally 70% of the expression of these genes and exons by RT-PCR. We isolated, within the set studied, a novel non-coding transcript that showed a highly specific pattern of expression by in situ hybridization in mouse embryos. Our analysis also shows that the genome presents gene trapping hotspots, which correspond to 383 known and 87 novel genes. These “hypertrapped” genes show minimal overlap with previously published expression profiles of ES cells; however, we demonstrate by real time PCR that “hypertrapped” genes are highly expressed in this cell type, letting us hypothesize that these genes could potentially contribute to the phenotype of ES cells. Thus the further studies of these genes, could help elucidate the “stemness” transcriptional profile. Although gene trapping was initially used as an insertional mutagenesis technique, our

study demonstrates its impact on the discovery of a substantial and unprecedented portion of the transcriptome.

In the second part of this work we focused our attention on conserved non-coding elements acting as enhancers. Generally speaking, non-coding regions are less conserved with respect to the protein-coding regions and their underlying syntax is not as clear. Conservation in non-coding sequence across the vertebrate subphylum has been shown to be a good predictor of regions which are involved in the regulation of the expression. Thus one way of predicting whether a DNA sequence is functionally important is the comparative analysis of orthologous non-coding regions across genomes belonging to this subphylum. In particular, in this work, using a global-local alignment on orthologous loci which takes into account the positional shuffling of regulatory regions across long evolutionary distances we identified over 20,000 vertebrate conserved elements, an order of magnitude more than previously reported. We demonstrated that 72% of these elements identified have indeed undergone to shuffling during the 450 million years separating fish from mammalian organisms. Furthermore we validated their function *in vivo* by testing their capability to act as enhancers when injected in zebrafish embryos and we demonstrated that more than 80% of these identified elements identified indeed act as enhancers often in a tissue specific manner.

Table of contents

| | |
|--|-----------|
| Acknowledgements | 5 |
| Abstract | 7 |
| Table of contents | 9 |
| List of Figures | 11 |
| List of Tables | 12 |
| Abbreviations | 13 |
| INTRODUCTION | 16 |
| Genomic features | 16 |
| The non-coding world | 18 |
| Repetitive sequences and mobile DNA sequences | 19 |
| Tandem repeats and Micro mini macro satellite repeats | 20 |
| Interspersed repeats: Transposable elements | 21 |
| Pseudogenes | 25 |
| Comparative genomics of non coding sequences | 26 |
| Verification of enhancer activity <i>in vivo</i> | 28 |
| Non-coding RNA genes | 29 |
| Functional genomics | 30 |
| Using the transcriptome to annotate the genome | 30 |
| How many genes? | 34 |
| Functional genomics: characterizing gene function | 35 |
| Gene trapping | 36 |
| Stem Cells as a model for transcriptome characterization | 38 |
| Aims of the thesis | 40 |
| MATERIALS AND METHODS | 51 |
| RNA extraction from ES cells | 51 |
| DNase digestion | 51 |
| cDNA transcription | 51 |
| Real Time quantitative PCR | 53 |
| Agarose gel electrophoresis | 54 |
| DNA sequence analysis | 54 |
| Isolation of DNA from agarose gels | 54 |
| Cloning of the PCR products | 54 |
| In vitro transcription | 55 |
| Precipitation of RNA | 56 |
| Quality control and quantification | 57 |
| Digoxigenin in situ hybridization | 57 |

| | |
|---|-----|
| Obtaining zebrafish embryos _____ | 58 |
| RESULTS _____ | 61 |
| Background _____ | 61 |
| Novel exons within known genes _____ | 62 |
| Identification of Novel Transcripts _____ | 65 |
| Expression profiling of a non-coding transcript. _____ | 67 |
| Trapping of genes correlates with their expression levels _____ | 68 |
| Verification of shuffled conserved elements in the vertebrate lineage _____ | 71 |
| Verifying SCE function _____ | 72 |
| DISCUSSION _____ | 87 |
| REFERENCES _____ | 94 |
| <i>Publications arising from this thesis</i> _____ | 104 |
| APPENDIX: PRIMERS USED _____ | 105 |

List of Figures

| | |
|--|----|
| Figure 1. Re-association kinetics. _____ | 42 |
| Figure 2. Fractions of different sequences in the human genome. _____ | 43 |
| Figure 3. Several transcript variants generated by different alternative splicing events. _____ | 44 |
| Figure 4. Schematic representation of a nested gene. _____ | 45 |
| Figure 5. Classes of interspersed repeats in the human genome. _____ | 46 |
| Figure 6. Different modes of transposition. _____ | 47 |
| Figure 7. Non-coding RNA transcripts. _____ | 48 |
| Figure 8. The basic trap vectors. _____ | 49 |
| Figure 9. Prediction of novel exons (1172) identified on RefSeq genes. _____ | 75 |
| Figure 10. Discovery of novel exons on known RefSeq genes. _____ | 77 |
| Figure 11. Prediction of 1,997 novel genes and 6,423 novel transcripts found within known gene loci. _____ | 78 |
| Figure 12. Discovery of novel genes based on trapclusters. _____ | 80 |
| Figure 13. In situ hybridization of trapcluster gene TCLG1417 on E14.5 mouse. _____ | 81 |
| Figure 14. Real-time RT-PCR verification of level of expression of hypertrapped genes. _____ | 83 |
| Figure 15. Shuffling categories of SCEs. _____ | 84 |
| Figure 16. Expression profiles of X-Gal stained embryos. _____ | 85 |
| Figure 17. Annotation of Trap Clusters. _____ | 93 |

List of Tables

| | |
|--|----|
| <i>Table 1. Verification of 40 novel exons tested by RT-PCR using RNA extract from ES cells, whole E14.5 embryo, heart, brain and eye.</i> | 76 |
| <i>Table 2. Results of RT-PCR verifications on ES cell RNA of 50 novel transcripts predicted to exist on the basis of gene trap sequence tags.</i> | 79 |
| <i>Table 3. List of the first 50 hypertrapped genes.</i> | 82 |
| <i>Table 4. Analysis of X-Gal staining in zebrafish embryos co-injected with the Hsp promoter and SCEs or control fragments.</i> | 86 |

Abbreviations

| | |
|-------------|--|
| Amp | ampicilline |
| bp | basepair |
| °C | degrees Celsius |
| cDNA | complementary DNA |
| CNC | conserved non-coding |
| CNE | conserved non-coding element |
| CNG | conserved non-genic |
| CNS(s) | conserved non-coding sequences |
| CNS | central nervous system |
| CpG islands | cytosinephosphatidiguanosine islands |
| DEPC | diethylpyrocarbonate |
| DNA | deoxyribonucleic acid |
| dNTP | deoxyribonucleotide triphosphate |
| DTT | Dithiothreitol |
| E14.5 | embryonic developmental day 14.5 |
| ECR(s) | evolutionarily conserved region(s) |
| EDTA | ethylenediaminetetraacetic acid |
| ENCODE | <u>E</u> ncyclopedia of <u>D</u> NA <u>e</u> lements |
| ES | embrionic stem |
| EST | expressed sequence tag |
| GADPH | glyceraldehydes-3-phosphate dehydrogenase |
| GECKO | genome-wide cell-based knockout |
| GSS | group support system |
| hES | human embrionic stem |
| hr(s) | hour(s) |
| kb | kilobase |
| LB | Luria Broth |
| LINE(s) | long interspersed element(s) |
| LTR(s) | long terminal repeat(s) |
| M | molar |
| MATRICES | Mouse Annotation Teleconference for RIKEN cDNA sequences |
| mg | milligrams |
| min | minute |
| miRNA | microRNA |
| MITE(s) | miniature inverted-repeat trasposable element(s) |
| ml | milliliter |
| mm | millimeter |

| | |
|----------|---|
| mM | millimolar |
| mRNA | messenger RNA |
| ncRNA | non-coding RNA |
| ng | nanogram |
| nm | nanometer |
| nM | nanomolar |
| nt(s) | nucleotide(s) |
| OD | optical density |
| ORF | open reading frame |
| PBS | phosphate buffered saline |
| PCR | polymerase chain reaction |
| PFA | paraformaldehyde |
| polyA | polyadenilation |
| RACE-PCR | rapid amplification of cDNA strands |
| RAGE | random activation of gene expression |
| rCNE | regionally-conserved element |
| RNA | ribonucleic acid |
| RNase | ribonuclease |
| rpm | revolution per minute |
| RT | room temeperature |
| RT-PCR | reverse transcriptase-polymerase chain reaction |
| SA | splice acceptor |
| SCE(s) | shuffled conserved element(s) |
| SINE(s) | short interspersed element(s) |
| SRP | signal recognition particle |
| TAE | Tris-acetate-EDTA |
| TCL | trapcluster |
| TCLG | trapcluster gene |
| Tris | 2-amino-2-(hydroxymethyl)propane-1,3-diol |
| TRP | transient receptor potential |
| TRPM | transient receptor potential membrane |
| TSS | transcription start site |
| TU | transcriptional unit |
| U | unit |
| UCE(s) | ultra conserved element(s) |
| UTR(s) | untranslated region(s) |
| UV | ultra violet |
| V | Volt |
| v | volume |
| w | weight |
| X-gal | 5-bromo-4-chloro-3-indolyl-b-D-galactosidase |
| ZFIN | zebrafish information network |
| µg | microgram |

| | |
|---------------|------------|
| μl | microliter |
| μM | micromolar |
| μm | micron |

INTRODUCTION

Genomic features

The eukaryotic genome contains several levels of complexity as demonstrated by re-association kinetics of its denatured DNA. In fact, DNA re-association occurs in three distinct phases, and each of them represents a different component (Fig. 1). Highly repetitive DNA represents 25% of the genome, DNA that is moderately repetitive represents a further 30% and 45% of the genome hosts non-repetitive DNA. The latter is known to contain coding genes as well as many other functional elements.

Functional elements of the genome can be classified further into coding and non-coding genes, pseudogenes, enhancers repressors and insulators, microRNA and many elements which probably still escape a complete understanding. Only a small portion of the genome, about the 2-3% of the mammalian genome, encodes mRNAs that encode for proteins, and the protein-coding sequence is located within large introns or intergenic regions (see Fig. 2).

The traditional genetic definition of a gene as a segment of DNA that is able to complement a mutant phenotype has become more complex in recent years, because it has become clear that the genomic sequence alone cannot be used to infer function, without taking into account a further complexity derived from alternative splicing. The set of transcripts that is derived from the genome composes the transcriptome. While in lower eukaryotic organisms the traditional paradigm of one gene, one transcript, one protein is likely to be valid for the majority of genes, in mammals it has become evident

in recent years that the transcriptome introduces a further, significant, layer of complexity.

The Fantom consortium (Carnici et al., 2005) has shown clearly that individual genomic loci can produce a multitude of overlapping transcripts. These transcripts, identified as full-length cDNAs, can be shown computationally to form clusters of overlapping sequences. A cluster of transcripts can arise from an expressed pseudogene, and an individual locus can encode clusters from both strands. This effort has shown clearly that the transcriptome is organized on the genome in complex regions, defined as transcriptional forests, which present a high complexity of sense and anti-sense, coding and non-coding transcripts.

Importantly, it has been shown that this variability is due to the fact that approximately 63% of the genome is transcribed at least from one strand (in comparison to previous findings that only the 2% of the genome is transcribed in protein coding mRNA), and that transcriptional units contains several alternative splice variants (Fig. 3), also due to the fact that many transcripts have multiple transcription start sites as well polyadenylation sites. Thus, overall, this project has clearly shown that the transcriptome is much larger and more complex than previously thought.

Coding sequences: exons

Protein-coding regions result from several coding sequences that are interrupted by stretches of non-coding sequences that are spliced out during mRNA maturation. Exons are defined as DNA sequences found in mature mRNA while introns are segments of DNA that are cut out in the final mRNA. Interestingly, some introns contain important information (such as splice enhancers and splice silencers, as well as enhancers) and

sometimes can even code for other completely different genes, the so called “nested genes” (see Fig. 4).

The non-coding world

For much over 50 years, the functional portion of the genome was considered to be the one that codes for proteins and, until recently, most evolutionary studies of DNA sequences have focused completely on this translated fraction. There are many theories on the origins of non-coding DNA which suggest that the bulk of these sequences is DNA debris with no meaning (Lynch et al., 2003) and invoke random accumulation of this “junk”, such as the action of selfish self-replicating elements (Orgel et al., 1980).

The idea that a wide proportion of the eukaryotic genomes contain elements conserved across evolution stems from the problem known as the “c-value paradox” where “c” stands for the total amount of DNA in the haploid genome. In fact, genome size does not correlate with organism complexity: for example, the unicellular organism *Amoeba dubia* contains approximately 200 times as much DNA as humans, while humans have about 7.5 times as much as the pufferfish *Fugu rubiprens*, although this organism has a comparable number of genes (Brenner et al., 1993). Most of the variation in genome size is due to the non-coding sequences, often very simple, repeated sequences.

The mammalian genome contains the instructions for many undiscovered non-protein coding RNA genes. About 0.5% of the human genome is represented by pseudogenes, but a large portion consists of introns and intergenic DNA. In fact, about half of the intergenic DNA consist of several type of transposons, while the remaninig non coding portion contains other elements responsible for the expression of genes,

structural elements responsible for chromosome function as well as remnants of evolution, all elements which constitute the non coding world or the so called “dark matter” of the genome (Hayashikizaki et al., 2006).

Despite recent elucidation of the extent of the genome which is transcribed, there are still large regions termed “gene deserts” which occupy a significant part of the genome. These are long regions which contain no transcribed sequences and without obvious biological function (Venter et al., 2001). Some studies have shown that gene deserts can contain regulatory sequences that act at a large distance to control the expression of neighboring genes (Nobrega et al., 2003; Kimura-Yoshida et al., 2004). These include *cis*-regulatory sequences that control gene expression (enhancers, insulators other boundary elements, and sequences that anchor genomic region to specific nuclear regions)(Dorsett et al., 1999; Bell et al., 2001; Carter et al., 2002) that can usually function in an orientation and often even position independent manner (Blakwood et al., 1998) influencing the activation or the specificity of a nearby promoter. In contrast, it has been demonstrated that some gene deserts do not seem to be essential to genome function since their deletion in mouse does not seem to cause a phenotype (Russel et al., 1982; Nobrega et al., 2004). Thus, further studies will be required to elucidate the role, if any, of these regions.

Repetitive sequences and mobile DNA sequences

The human genome contains also stretches of repeated non-coding elements of various length and copy of number (identical and/or similar copies). Repetitive sequences make up at least 50% of the entire human genome. They are classified by function and dispersal pattern. These repetitive sequences are called “tandem repeats” if present as a sequence motifs lying adjacent to each other in the same block, or “interspersed repeats” if the repetitive sequences are scattered along the genome as single units flanked by

unique sequence. Although their relevance as functional elements is still unclear, even if we assumed that repeated elements do not play an important functional role, a large amount of non-coding non-repetitive DNA remains to be elucidated.

Tandem repeats and Micro mini macro satellite repeats

Tandem repeats contain successive identical repeat units. This class of elements includes satellite DNA, minisatellite and microsatellite repeats; satellite sequences are quite variable in repeat size and in array size. Microsatellites are the smallest, at a repeat size of 4 bp or less. Moreover, recently macrosatellites have been discovered which are moderately repetitive and contain tandem repeats of a larger size in some cases ORFs can be as long as 4-10kb long (Gondo et al., 1998).

Whether these sequences often as short sequences as 2-3 bps and repeated as often as thousands of times, play a functional role is still unclear. Often these sequences appear to function collectively rather than individually, and their dispensability is not an indicator of non-functionality. It is noteworthy that in the genome of *Fugu rubripes*, a highly compact vertebrate genome most repeat families found in other vertebrates are present, although in very limited copy number, sometimes as small as a single copy (Aparicio et al., 2002). Satellite DNA sequences are abundant in constitutive heterocromatine. In particular they are involved preponderantly in the organization of the centromeres, the sites in which every chromosome attach to cellular tethers and are pulled during mitosis. Moreover, minisatellites are enriched in subtelomeric regions of the chromosome.

Interspersed repeats: Transposable elements

Retroelements

One of the most common classes of repeats (~35% of the genome) is that of dispersed retroelements (Jurka et al., 1998). Retrotransposons can be classified into two categories: autonomous and non-autonomous (Fig. 5). While the former encode for a protein necessary for transposition, the latter do not encode a protein. For this reason the latter need a separate protein product encoded by another transposon to perform their transposition. Another classification of these transposable elements is based on the mode of transposition (Finegan et al., 1989). The “class I mobile elements” is capable to reproduce itself using an RNA intermediate which is reverse transcribed to DNA by a reverse transcriptase enzyme encoded on intact elements (Fig. 6). It has been observed that these elements require an RNA polymerase (II or III) to be transcribed into RNA and thereafter be transposed, while the original DNA copy is preserved in the same location. Short and long interspersed elements, named SINEs and LINEs respectively, represent the majority of this class of repeats and they form a group called non-LTR elements. The remaining part of this class comprises LTR transposons, structurally similar to integrated retroviruses, and retrogenes. Finally the elements belonging to class II move by a conservative “cut and paste mechanism”, which involves the excision of the donor element is followed by its insertions elsewhere in the genome (Fig. 6).

LTR retrotransposons

LTR retrotransposons are remnants of endogenous retroviruses which represent 8% of the genome and are usually 7-9kb long. They contain, like the proviruses, long terminal repeats (LTR), gag, pol, and prt genes with the difference that one of the

proteins responsible for the infection, the env protein, is mutated or missing. Thus, these elements can only move within cells. The human genome contains only “evolutionary fossils” of these elements which are highly mutated and are not capable of transposition any longer.

LINEs

LINEs (long interspersed nuclear elements) are autonomous retrotransposons. These sequences represent 21% of the human genome. In particular the most abundant in humans are Alu and the LINE-1 sequences (Lander et al., 2001). LINE-1 sequences alone comprises the 17% of the genome. The basic active element, about 6 kb long, called L1 contains two open reading frames, ORF1 and ORF2, a 5'UTR, which acts also as a promoter and a 3'UTR containing a polyadenylation signal. It is known that ORF2 is responsible for integration in the genome and that it contains an endonuclease domain as well as a reverse transcriptase domain. The function of the product of ORF1 is still unclear, it is only known that it binds to L1 mRNA. After the L1 mRNA transcription, it is transported in the cytoplasm, thus ORF1 is translated. The translation, then, is restarted to an internal ribosome entry site to translate ORF2. This process in eukaryotes occurs rarely so that only a little portion of L1 has its ORF2. Both proteins binds L1 and this complex is traslocated into the nucleus. The ORF2 acts by cutting the DNA at the target site. This process is not particularly specific, but occurs preferentially for AT rich sequences. This cut occurs unequally and generates sticky ends; thus the free 3'OH group is used by the reverse transcriptase encoded by ORF2 for the synthesis of the first cDNA strand. The mechanism of synthesis of the second cDNA strand is still unknown, but it is known that the end result is a stable integration of a double stranded L1 DNA in a new

location within the genome. Thus LINEs can be considered vectors for DNA shuffling thus contributing to DNA relocation events of small fragments (such as exons and enhancers) within the genome. The L1 element is flanked by target sites for duplication which span 7-20 bps. Owing to the fact that the reverse transcriptase does not always finish transcription of the first strand, the newly formed copy is often truncated at the 5' end. Moreover, the lacking of the proofreading activity in the process leads to the introduction of several mutations within the new copy.

SINEs

SINEs (short interspersed nuclear elements) do not encode for any protein and typically their length is shorter than 500bp. Among them the Alu elements (which derive their name derived from the identification of AluI restriction sites) represent about 11% of the human genome. These elements share a consensus of about 282bp that derives presumably from the SRP (signal recognition particle) RNA subunit (7SL RNA). Alus are transcribed by the RNA Pol III, the same enzyme which transcribes the 7SL RNA gene. Moreover, Alus are capable to bind two SRP proteins (SRP9 and SRP14). It has been suggested, therefore, that Alus can bind to the ribosome machinery and that through their polyA tails they might bind nascent ORF2 proteins from LINE1 RNAs and force these proteins to induce the reverse transcription and integration of their RNA rather than LINE-1 mRNA.

Repeats may be also be responsible for epigenetic control mechanisms, or other modifications of gene activity, based on modifications of the DNA itself rather than its sequence. It has been hypothesized, for example, that a repeat-induced process involving L1 retroelements might be responsible for the X-inactivation, a process necessary to

maintain proper gene dosage in females who have two X-chromosomes (Neumann et al., 1995).

It can happen very rarely that a conventional cellular mRNA is subjected to reverse transcription and transposition from an enzyme deriving from L1 or other retrotransposons in which case the gene undergoes duplication. The new copy of the gene in this case will lack of its promoter region as well as its introns, and thus in most cases will lose its function, becoming a “processed pseudogene”. Processed pseudogenes are distinct from “ordinary” pseudogenes which instead derive from a duplication event of a whole genomic portion, and thus maintain their original gene structure comprising exons, introns, promoters and so on. Sometimes insertions can occur during this process disrupting the original function of the gene and causing a genetic disease.

Elements encoding Transposase

These elements belong to class II and comprise inverted repeats (10-500 bp) at their termini and encode trasposase which catalyzes trasposition. Following excision they shift elsewhere in the genome where they insert by a non-replicative mechanism. It has been shown that the human genome contains sequences originated from more than 60 different DNA transposons.

MITEs

MITEs (miniature inverted-repeat transposable elements) constitute another group of mobile elements (Feshotte et al., 2002). They have short terminal inverted repeats, and their length is comprised between 125-500 bp. They were firstly identified in plants, and subsequently observed in mosquito, zebrafish and human (Dufresne et al., 2007). Their mechanism of transposition is still unknown, but they do not appear to be autonomous.

MITEs appear to be preferentially associated with genes and thus might play a significant role in generating genetic variation (Dufresne et al., 2007).

Effects of repetitive elements on gene expression

Mobile elements and repetitive elements can alter the structure of the genome and can regulate gene expression of the genome in several ways. Firstly, as previously described, transposition may disrupt functional genes. Many transposable elements have a constitutive promoter that can drive an inappropriately expression of a gene downstream. On the other hand if the promoter of the transposable element is opposite with respect to a neighbouring gene, it can initiate transcription of an RNA transcript which is complementary to the gene mRNA, and thus disrupt the endogenous expression of the gene via antisense RNA mediated silencing.

Pseudogenes

Pseudogenes belong to the set of non-coding transcripts which are less likely to have a biological role (Cheng et al., 2005; Carnici et al., 2005). The cDNA collection obtained by FANTOM3 contains several transcripts that seem to encode for proteins, but which contain a few mutations disrupting the ORF, which could be considered pseudogenes. The definition of pseudogenes has been modified over time. Initially pseudogenes were considered genomic sequences which resemble functional genes, but which for some reason have been inactivated. As noted earlier, some derive from the insertion of mobile elements within open reading frames (ORFs) of functional genes, while others are the result of “processed genes”, i.e. the sequence indicates that probably a retrotranscription event has taken place (with RNA being used as a template to make DNA) and has resulted in the re-integration of the generated DNA within the genome.

While the common view is that most pseudogenes do not perform a clear biological function but are, rather, evolutionary fossils, recent findings indicate that some are clearly functional (Hirotsune et al., 2003; Zheng et al., 2005), by binding transcription factors and impeding them from being involved in the activation of gene expression.

Comparative genomics of non coding sequences

One of the aims of genomics is try to understand how genomes are organized and in particular which sequences are involved in the complex mechanisms involved in the regulation of gene expression. The sequencing of a large number of genomes, in particular within the chordate subphylum, has lead to the utilization of comparative genomics techniques. The basic principle of comparative genomics is that of identifying portions of the genome whose sequence has changed significantly less than expected during evolution, indicating potential functional constraints and thus enabling us to distinguish potentially functional non-coding DNA from junk DNA. Generally speaking non-coding regions are less conserved than protein-coding genes. However, the first large scale comparative genomics analysis, which was done when the first draft of the mouse genome had become available, showed clearly that protein-coding sequences only account for approximately a fifth of the total amount the genome which is subject to purifying selection (International Mouse Genome Consortium, 2002), thus implying that a relatively large amount of non-coding DNA is likely to be functional. These segments of highly conserved elements are usually embedded among large dissimilar segments producing a mosaic picture of genomic conservation.

Studies of small genomic regions had demonstrated the possibility to identify putative genes as well as regulatory elements looking at cross-species conservation already prior

to the first drafts of entire genomes (O'Brien et al., 1999; Ansari-Lari et al., 1998). Comparative analysis between mouse and human genomes suggested that 5% of genomic DNA is under active selection which is likely to be associated with a functional role (Waterson et al., 2002). Many of these conserved regions correspond to protein coding exons, while the remaining sequences, generally called “conserved non-genic sequences” (CNG) or “conserved non-coding sequences”(CNS) (Hardison et al., 2000) seem to be involved in important regulatory activities (Dermitzakis et al., 2006). The latter constitute a significant portion of non-coding DNA and have become the focus of deeper investigations recently. Intriguingly it has been shown that CNSs represent only a subset of regulatory elements and at the same time only a subset of them are regulatory elements (Nobrega et al., 2003). In fact, only a fraction of these sequences can be associated with transcriptional regulation, such as enhancers (Nobrega et al., 2003; Bejerano et al., 2004), while it is not clear whether the rest of them bear a biological function. Supporting the notion that not all of these are directly related to the regulation of transcription of specific genes, it has also been observed that they are scattered along the genome independently of gene density (Dermitzakis et al., 2004; Dermitzakis et al., 2005).

Evolutionarily conserved regions (ECRs) are found both in coding and non-coding regions and have been identified computationally comparing two mammalian genomes such as mouse and human and using a window of length 70-100 bp and a threshold of percentage identity ranging from 70% conservation (Loots et al., 2000; Dermitzakis et al., 2003) to complete identity. Comparisons of the human genome against the genomes of distantly related vertebrates, moreover, have revealed an abundance of highly conserved non-coding elements (CNEs) (Boffelli et al., 2005; McEwen et al., 2006). Interestingly, a

property of human CNEs is that they cluster in genomic regions containing transcription factors and genes involved in the regulation of development (Bejerano et al., 2004; Woolfe et al., 2005; Vavuri et al., 2006).

Although non coding sequences generally lack sequence conservation among divergent species (Thomas et al., 2003), comparisons between human and the Japanese pufferfish (*Fugu rubripens*) show that those non-coding elements which do present significant sequence conservation often also play a role *in vivo* (Marshall et al., 1994; Rowich et al., 1998; Kammandel et al., 1999; Bagheri-Fam et al., 2001; Ghanem et al., 2003; Lettice et al., 2003; Nobrega et al., 2003; Santagati et al., 2003; Spitz et al., 2003; Kimura-Yoshida et al., 2004). The common ancestor shared by both *Fugu* and human lived about 450 million years ago (Kumar et al., 1998), implying that sequences which show significant conservation between these two species, including non-coding sequences, are highly likely to play a role in vertebrate life.

Verification of enhancer activity *in vivo*

A general strategy to test whether non-coding regulatory sequences are functionally relevant involves assaying their ability to up-regulate (or down-regulate) reporter genes *in vivo*. “Enhancer” assays using transgenic animals, especially in the case of transgenic mice, are very slow, costly and laborious, but have so far been one of the main sources of data on the function of non-coding DNA around, in particular for developmental genes (Nobrega et al., 2003, Pennacchio et al., 2006). In recent years an alternative approach has emerged, which has proven to be very useful to tackle this issue, which uses transient expression assays in zebrafish (*Brachidanio rerio*) embryos obtained by co-injection the candidate enhancer sequence with a promoter/reporter construct in the fish (Muller et al., 1997; Muller et al., 1999; Dikmeis et al., 2004). The zebrafish model presents on the one

hand a divergent genome suitable, although challenging, for comparative genomics analysis and on the other hand it is an extremely tractable experimental system. The experimental advantages are represented by the fact that a large number of fertilized eggs are available and easily modified by micro-injection, that the developing embryos are transparent and contain many easily identifiable cells, and finally that the detailed anatomical, physiological and developmental properties are known for many of these cells. Although the pattern obtained in this transient expression assays are mosaic, it is feasible to screen hundreds of individuals embryos at the same time, thus collating mosaic patterns into a final compound image.

Non-coding RNA genes

For several years molecular biology was based on the central dogma that stated that genetic information stored in DNA is transferred into RNA through transcription and is then finally decoded by translation of RNA into proteins. In this view RNA molecules played a passive role of mere messengers. Today it is well established that RNAs can play much more active roles within a cell and that several classes of RNA molecules exist which serve a function without encoding a protein message (Fig 7). RNAs can be thus divided into two main classes: messenger RNAs which are destined to be translated into proteins, and non-coding RNA (ncRNA), many of which are not well characterized yet, but which can be broadly classified as such because they generally do not encode for a protein. Non-coding RNA (ncRNA) transcripts are can play a multitude of roles, have their own structure and act as regulatory and/or catalytic molecules. Although the FANTOM3 project estimated that at least about 28,000 ncRNAs exist in mouse (Liu et al., 2006) the total number of ncRNA genes present in the mammalian genome is far

from clear, let alone their function. It has been observed that a large portion of ncRNA transcripts have introns (Ota et al., 2004), which raises the possibility that the primary transcript could be inactive and the subsequent cleavage and splicing maybe required to generate an active RNA molecule. The nature of these molecules is quite variable (small or multicopy), and their conservation across genomes is rather poor, thus it is complex to detect them and annotate them appropriately. The size of ncRNA molecules is also extremely variable from some as small as 22-25 nucleotides (which is the case for miRNAs) to thousands of nucleotides (such as ncRNA involved in silencing) (Hutvagner et al., 2002). The processes in which they have been shown to be involved are wide, from transcriptional and post-translational regulation, to chromosome replication, mRNA stability, protein degradation and so on (Hutvagner et al., 2002; Brandl et al., 2002). Thus it is an entire new world, likely to be at least as complex as that of proteins, which awaits to be discovered.

Functional genomics

Using the transcriptome to annotate the genome

Once the sequence of several mammalian genomes was completed until annotation tasks focused on the annotation of genes within the sequence, initially relying on mapping protein and cDNA sequences of known genes (which had been cloned in the past 50 years individually) and cDNA or EST sequences, as well as any other genes which could be predicted either by sequence similarity (for orthologs and paralogs in particular) or by *ab initio* gene prediction, which is based largely on the basic properties of coding genes (such as 3rd codon position degeneracy, ORF detection and hexamer statistics). As genome annotation developed so did genome browsers such as Ensembl

(<http://www.ensembl.org/>), the UCSC genome browser at the University of California at Santa Cruz ([http:// genome.ucsc.edu](http://genome.ucsc.edu)), as well as the browser present at the National Center for Biothechnology information ([http://: www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). These browsers are user friendly and allow users to scroll along the chromosome and zoom in or out to any scale, and display information at several levels of detail.

Although these initial approaches were incredibly useful to provide a first annotation “map” of the genome (and they are still valid and utilized now) it quickly became apparent that a good annotation was heavily reliant on the datasets that were used to produce it and thus strong efforts were put in place to produce larger and more diverse datasets exploring the full functional potential of the genome.

One of the widely tackled issue in an attempt to provide deeper functional annotation of the genome was that of characterizing comprehensively the transcriptome. The first approach which had a deep impact in this sense was the high throughput sequencing of cDNA ends (ESTs). The UniGene project (<http://www.ncbi.nlm.nih.gov/UniGene>) for example, assembled into clusters all available EST sequences creating a public database which, on the one hand was integrated in genome annotation pipelines, and on the other hand became a resource in its own right which can provide information, for example, on the relative tissue distribution of each cluster, yielding hints on the potential expression of a novel gene. An inherent disadvantage of this method is represented by the fact that while abundant transcripts have been sequenced thousands of times, many rare transcripts are completely absent from these EST databases, in particular those which are expressed only in very specific cell types, and are therefore very rare in whole tissue/organ libraries.

A similar approach has been developed based on the isolation and the sequencing of full-length cDNAs. The RIKEN Mouse Gene Encyclopedia Project, amongst others, has adopted this approach in a systematic manner providing a comprehensive dataset for the eukaryotic transcriptome. The RIKEN group used several complementary techniques to produce full-length cDNAs (Carnici et al., 2003). These techniques required (1) a novel reverse transcriptase reaction, (2) a novel 5'capture technology, (3) novel approaches to normalize and subtract cDNA libraries. Furthermore in order to fully annotate all the collected cDNAs, as well as to perform in-depth follow-up studies on the dataset an international consortium called FANTOM was put together. Initially, the consortium produced the FANTOM1 collection comprising about 21.076 cDNAs, and developed a simple web-based annotation interface for this dataset (Kawai et al., 2001). Already within the first collection it was observed that there was some redundancy in the set of cDNAs obtained (i.e. some clones were picked with different, but overlapping sequences). One of the causes identified was the high level of 3' end variation (due to alternative polyadenylation/termination signals) in mammalian mRNAs. In this first round of the project a large number of clones remained "unclassified" because their annotation was not very clear at the time (due to the lack of an ORF, etc). The project was thus extended to shed further light on the data obtained. During Fantom2 an interface was created that became an all-online annotation system from remote sites via the Internet, through the "Mouse Annotation Teleconference for RIKEN cDNAs Sequences" (MATRICS). In this way, the knowledge of the mouse transcriptome was considerably extended, however the cDNAs collected still covered only half of all the genes predicted in the genome. Finally the collection was expanded utilizing a much larger number of

tissues and cell lines as RNA sources in the third round of the project, Fantom3. Fantom3 was a major turning point for the project, as it became apparent that approximately half of the genome is transcribed into non coding transcripts, and that the genome is organized into transcriptional forests (TFs) comprising a multitude of coding and non-coding, sense and antisense transcriptional units (TU) and transcriptional deserts, which lack any evidence of transcription (Carnici et al, 2005).

Interestingly, FANTOM3 also deployed several techniques complementary to the mere identification of full-length cDNAs, such as CAGE, a technique aimed at obtaining the first 20 nucleotides of all transcripts screened, which are then concatenated, much like in SAGE, and sequenced, thus enabling fast and cheap sequencing of a very large number of transcription start sites, yielding novel information on the usage and frequency of transcription start sites in the mammalian transcriptome. This and other complementary techniques used in Fantom3 clearly demonstrated that both transcriptional start sites (TSSs) and transcriptional termination sites exceeded the number of Transcriptional Units identified, thus underlining that the usage of alternative start and termination sites is yet another form of complexity embedded in the mammalian genome, despite the fact that the number of genes is that is very similar across vertebrates.

The data obtained in recent years on transcriptional start sites allowed the development of novel algorithms aimed at transcription start site detection, such as EPONINE, a program which aims to predict the exact location of the transcriptional start site (TSSs) (Down et al., 2002) for a subset of genes. The TSS model utilized corresponds to the observation that promoters are often associated to CpG islands, as well

as specific motifs such as the TATAAA motif tightly distributed at position -30 relative to the transcription start site.

As described earlier, it is now clear that a large part of the genome bears functional elements that escape the well-known rules of protein coding genes and often also those of transcription as a whole. Thus a recent project was developed to tackle this specific question, the ENCODE (Encyclopedia of DNA Elements) Pilot Project (The ENCODE Project Consortium 2004, 2007). The aim of the project is the mapping of all the varieties of features present in the genome, such as genes, promoters, enhancers, silencers or repressors, exons, replication origins and termination sites, as well as chromatin modifications, methylation sites, conserved sequences, etc. The ENCODE project provided the identification of novel TSSs as well as the arrangement of regulatory sequences and binding sites for transcription factors around TSSs (Denoeud et al., 2007, Trinklein et al., 2007; Xi et al., 2007; Zhang et al., 2007).

How many genes?

Although the sequence of the human genome can be considered virtually complete, several debates have developed on the definitive catalogue of the genes that it contains. In fact, the rapid completion and the public release of the mouse and human genomes has led to a decrease of the number of genes predicted in the mammalian genome (Waterson et al., 2002). The Human Genome Sequencing Consortium estimates that the actual number of human genes is comprised between 20,000 and 25,000, strikingly lower than the early estimates of far more than 30,000 (Lander et al., 2001). For a long time the total number of genes has been a matter of debate; early estimates ranged from 28,000 to 120,000 genes, based on expressed sequence tag (EST) clustering (Ewing et al., 2000; Liang et al., 2000). Today, thanks in part also to the Fantom3 project these large

discrepancies have been understood to arise from the distinction between loci (which are approximately 20,000) and the transcriptional forests contained within them (which contain over 100,000 transcripts), and the overall complexity attributed to coding and non-coding transcripts, alternative splicing, alternative transcription start sites, etc. Before a large number of full-length cDNA sequences became available, each appeared as a distinct entity, rather than as a part of a complex transcriptional forest, thus impacting erroneously gene counts.

Functional genomics: characterizing gene function

The massive increase in sequencing projects allowed to rapidly expand the realm of both cDNA and genomic DNA information. It quickly created a gap, however, between the rapid discovery of genes and the slow process of actually identifying their function both within a physiological as well as, importantly, a pathological context. Thus, it became crucial to tackle this information gap and, to this end, scientists developed novel functional genomics strategies to develop experiments designed to discover and characterize the function of novel genes in a reasonably high throughput manner.

Several strategies were thus devised to identify, isolate and characterize genes by disrupting their function and observing the phenotypes induced. Some of the techniques developed are enhancer and promoter traps (Friedrich et al, 1991), gene traps, random activation of gene expression experiments (RAGE) as well as genome-wide cell-based knockout (GECKO). Finally, owing to gene-targeting techniques, transgenic mice have also proven crucial for the understanding and evaluation of gene function as well as to develop models of human disease based on specific single or multiple gene knock-outs.

Gene trapping

High-throughput gene trapping is a random approach for inducing insertional mutations within the genome; in recent years, this technique has become very important to study development exploiting the use of embryonic stem (ES) cells *in vitro* and *in vivo*. The principle behind gene trapping is essentially the random insertion of a DNA vector designed so that if the insertion happens within an existing gene locus its activation is detected via a reporter gene embedded in the vector. Gene trap vectors simultaneously inactivate and report the expression of the trapped gene at the insertion site, as well as providing a DNA tag for the rapid identification of the disrupted gene.

Three main types of entrapment vectors have been described: (1) enhancer trap vectors, which have to be integrated near an endogenous enhancer in order to activate the reporter gene that is fused to a minimal promoter (Bellen et al., 1996) (2) gene trap vectors, which need to integrate within an already actively transcribed gene in order to work and (3) promoter trap vectors which also need to be integrated within an existing gene, but since the vector bears also a promoter, the gene does not necessarily need to be active. The principal element of all gene trapping vectors is a gene trapping cassette consisting of a reporter gene and/or a selectable marker gene flanked by an upstream 3' splice site (splice acceptor (SA)) and a downstream transcriptional termination sequence (polyadenylation sequence (polyA)). When inserted within the intron of an expressed gene, the gene trap cassette is transcribed from the endogenous promoter in the form of a fusion transcript in which the exon(s) upstream of the insertion site is spliced with the reporter/selectable marker gene. Since transcription is terminated prematurely at the polyadenylation site contained within the vector, the processed fusion transcript encodes a truncated and non-functional version of the cellular protein as well as the

reporter/selectable marker (Stanford et al, 2001). When gene traps are introduced into ES cells, they integrate more or less randomly across the genome, although some preferential trapping events are known to occur (Durick et al, 1999; Skarnes et al, 1992; Skarnes et al, 1995; Von Melchner et al, 1992). Antibiotic resistant ES cell colonies are easily selected and expanded in vitro, and clonal cells can be isolated for injection into mouse blastocysts or differentiation in vitro. Expression of the gene trap is assayed for reporter gene expression (e.g. β -galactosidase activity), and staining is indicative of an insertion event. The transgene is only activated when it integrates correctly within a transcriptional unit; however it is known that some translation fusions (frameshifts) inactivate the reporter activity or may target the translated proteins into subcellular location where reporter activity is not easily detectable.

The possibility of developing mice derived from these “trapped” ES cell lines have permitted the identification of many novel developmentally regulated genes with specific spatio-temporal expression patterns as well as a better characterization of known genes. By selecting for the activation of the reporter gene in cell culture, the rate of gene disruption in recovered clones approaches 100%, and the random insertion of exogenous DNA into single sites in mammalian genome (gene trapping) provides a genome-wide strategy for functional genomics. ES cell cultures thus provide a simple model system for studying the genetic pathways that regulate embryonic tissue development and permit high-through-put screening of clones for tissue restricted gene trap expression (Bonaldo et al, 1998).

Stem Cells as a model for transcriptome characterization

Embryonic stem cells can be maintained in culture as totipotent cells (i.e. cells that can give rise to all type of cell lineages) under appropriate growth conditions and can be easily genetically altered. ES cells are one of the richest sources of transcriptional diversity; in fact they are known to express (at low levels) the majority (60%) of known genes, probably in relation to their pluripotent state, as though many genes were ready to be upregulated depending on which differentiation stage is undertaken. On the other hand these cells are also known to have a set of genes expressed at significant levels which are likely to be responsible for their “stemness” phenotype. The recent advances made by *Fantom3*, furthermore suggest that probably many other unique transcripts, either entirely novel, or derived from splice variants of known ones, are likely to be also involved and probably remain to be identified.

It is worth noting also that although the sequence tags obtained from trapping experiments in ES cells are similar in quality to ESTs (i.e. short, single pass, low quality reads of sequence in most cases much shorter than the transcript they are derived from), they are quite different in nature. ESTs derive from cDNAs and have therefore specific biases attached to the method utilized to obtain them. For example they are usually only 5' and 3' ends of full-length transcripts, and their detection is highly dependent on transcriptional levels. Sequence tags derived from gene trapping experiments are only in part dependent on transcriptional levels (since some vector are able to trap genes that are not expressed in ES cells), and derive directly from a genomic integration of the vector. Interestingly, several preferred integration sites, or “hot spots” have been observed (Hansen et al., 2003). Moreover, it has been demonstrated that these gene trap hot spots

are not sequence specific and are not related to gene size, suggesting that they are defined by secondary chromatin structure (Hansen et al., 2003).

Bioinformatics-based approaches have accelerated the evaluation of mutant clones (originated by gene trap, RAGE and GECKO experiments) leading to the rapid identification of informative cell lines on an unprecedented scale. The combination of this resource with other large-scale approaches including bioinformatics, expression profiling and in situ hybridizations just to name a few, is a powerful tool which enables to quickly provide some hypothesis with regards to the specific biological process or disease state with which a novel gene might be associated, thus providing a clue for further testing. For example, a sequence tagged gene-trap library of > 270,000 mouse ES cell clones has recently been developed and has been employed together with a functional screen of knock-out mice to identify genes regulating blood pressure (Friedel et al., 2005) (e.g., <http://baygenomics.ucsf.edu/overview/welcome.html>). Efforts are also underway to make ES cell lines with gene traps freely available for researchers so that transgenic mice containing a potential gene of interest can be made to further understand the role of specific genes in development and disease (Skarnes et al., 2004).

Aims of the thesis

The completion of the sequencing and annotation of the mouse genome (Waterson et al., 2002) suggested that the understanding of the number and the function of most genes in the genome would be accomplished swiftly. Recently the FANTOM Consortium has demonstrated quite evidently that the annotation of the genome is far from being completed. Quite on the contrary Fantom has demonstrated that the genome is organized into transcriptional forests, that present a complex array of sense and anti-sense, coding and non-coding transcripts (Carnici et al., 2005) and that we only begin to understand the multi-dimensional complexity which is overlaid on the mono-dimensional layer of DNA sequence.

In our study we have used data derived from a gene trapping approach in mouse ES cells to re-annotate the mouse genome as well as shed light on gene trapping hot spots. Stem cells express a large number of transcripts at low levels, which are “dormant” ready to be activated upon differentiation. They also express a set of genes, some of which have still unknown function, at significant levels, likely to be involved in maintaining the “stemness” state (Boheler et al., 2003). Although gene trapping is not a novel resource, it has not been used extensively in the context of genome annotation, and with this work we demonstrate that it is indeed a very significant source of data to identify novel features of the mouse genome as well as to characterize further the genes involved in the “stemness” phenotype.

In the second part of this work we investigated the function of specific non coding elements that shows a conservation across divergent organism. We found that the majority of these elements undergo shuffling across evolution (thus they were called

shuffled conserved elements, SCEs) and we prove that the majority of them act successfully as enhancers *in vivo*.

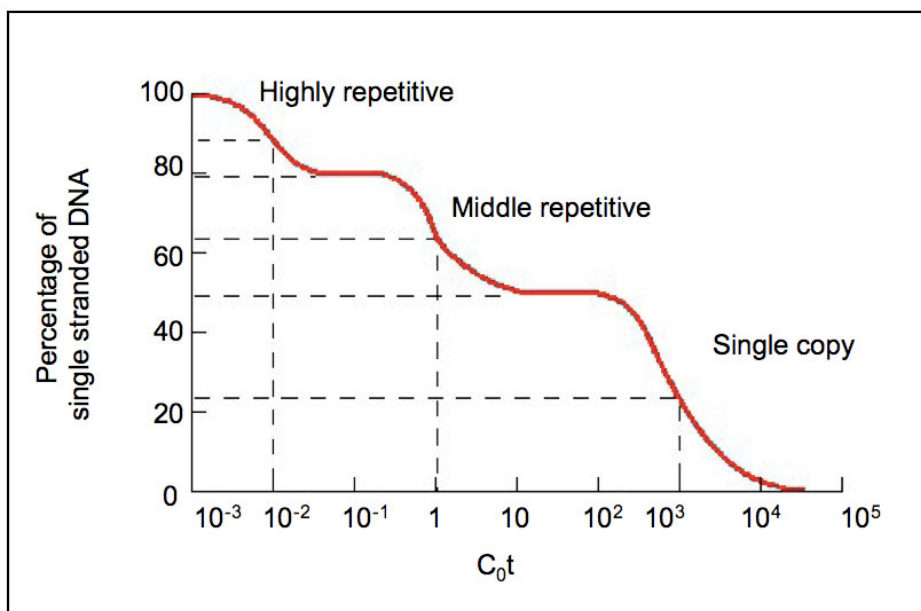


Figure 1. Re-association kinetics.

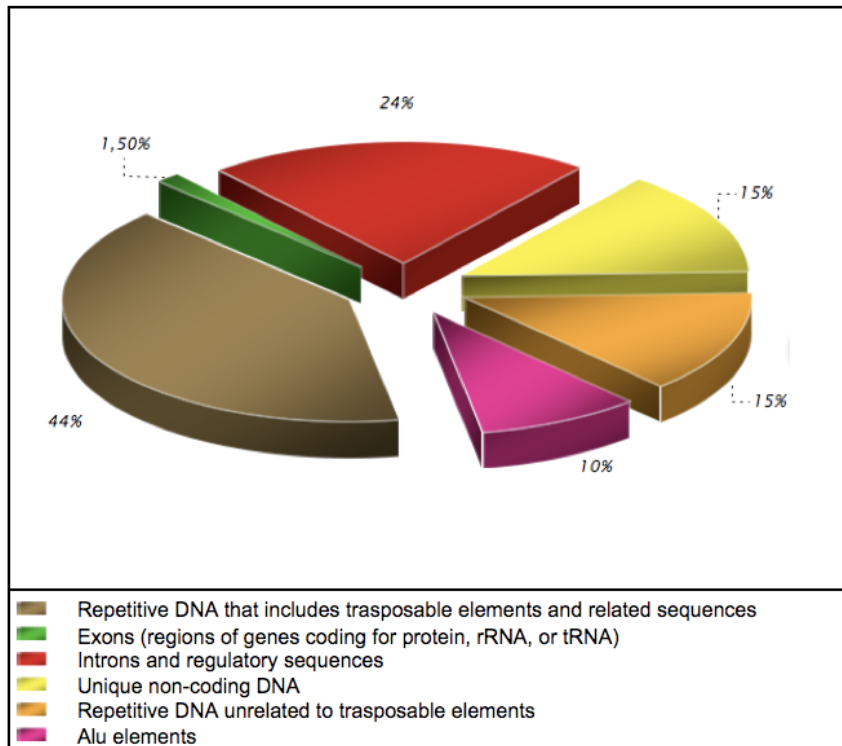


Figure 2. Fractions of different sequences in the human genome.

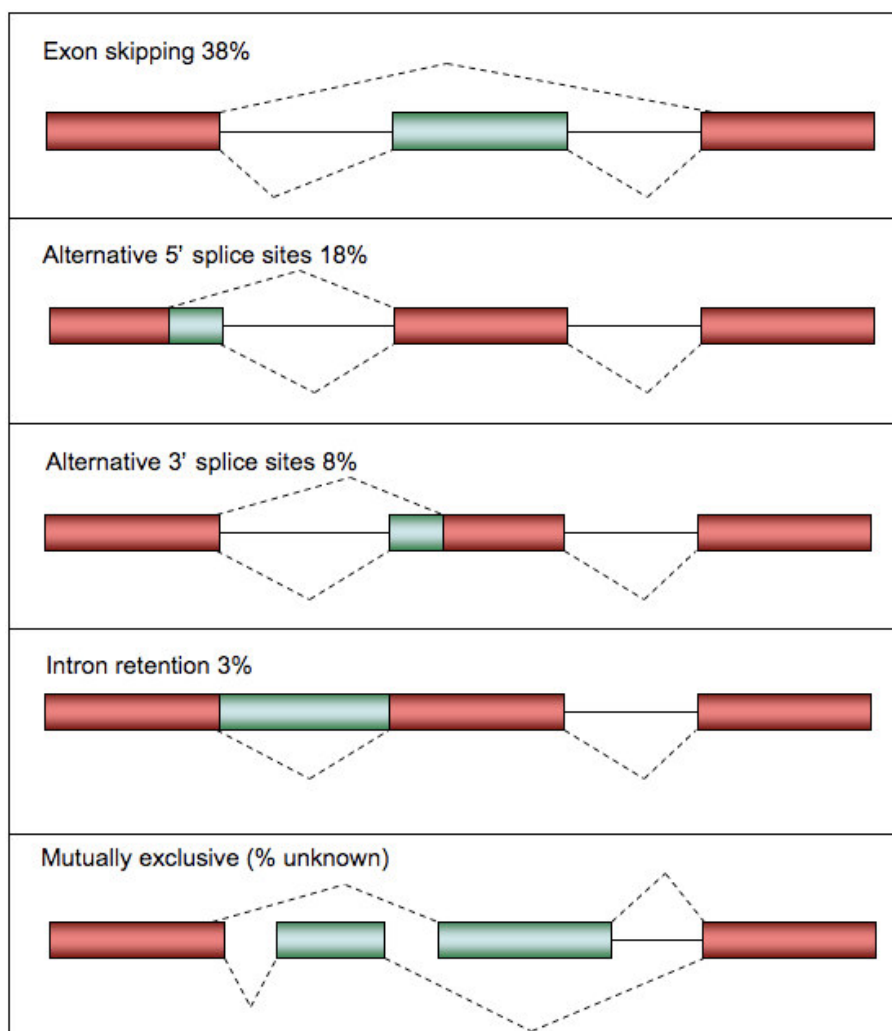


Figure 3. Several transcript variants generated by different alternative splicing events.

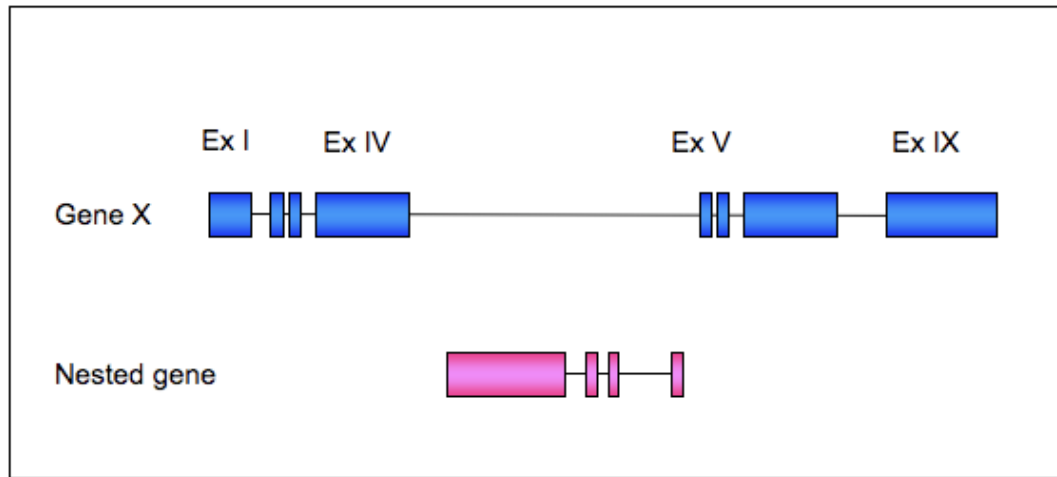


Figure 4. Schematic representation of a nested gene.

The gene is in the same direction of the gene X but is completely contained within its introns.

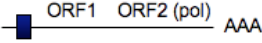

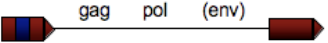
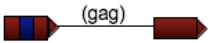
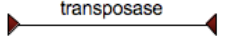

| | | | Length | Copy number | Fraction of genome |
|--------------------------|-----------------|---|-------------|-------------|--------------------|
| LINEs | Autonomous |  | 6-8 kb | 850,000 | 21% |
| SINEs | Non- autonomous |  | 100-300 bp | 1,500,000 | 13% |
| Retrovirus-like elements | Autonomous |  | 6-11 kb | 450,000 | 8% |
| | Non- autonomous |  | 1.5-3 kb | | |
| DNA transposons fossils | Autonomous |  | 2-3 kb | 300,000 | 3% |
| | Non- autonomous |  | 80-3,000 bp | | |

Figure 5. Classes of interspersed repeats in the human genome.

Blue rectangle: promoter; red block: LTR (long terminal repeat); triangle: short terminal repeat.

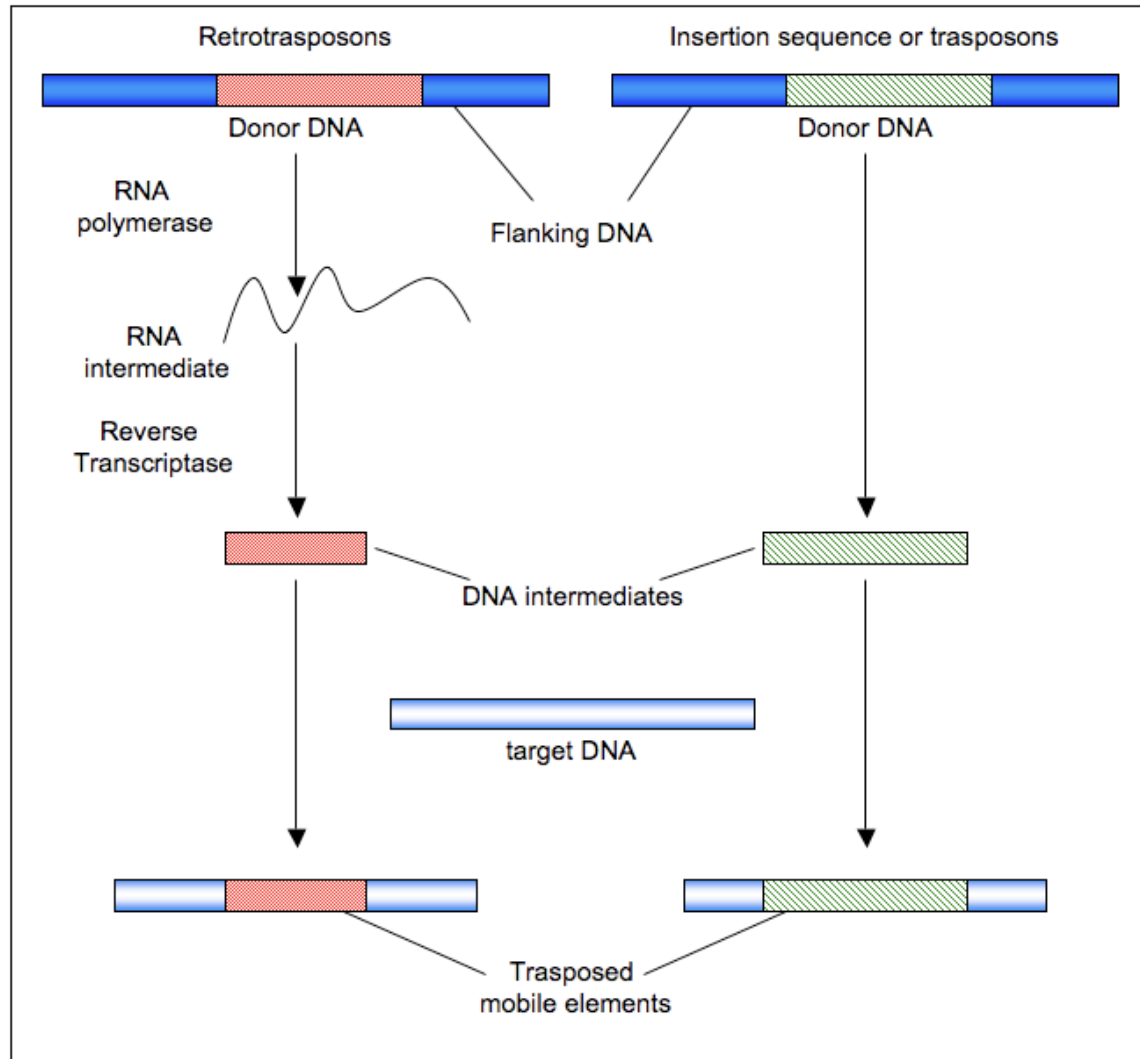


Figure 6. Different modes of transposition.

The class I mobile elements (non-LTR elements) uses an RNA intermediate to reproduce itself (shown on the left panel). The class II mobile elements move by a conservative “cut and paste” mechanism (shown on the right panel).

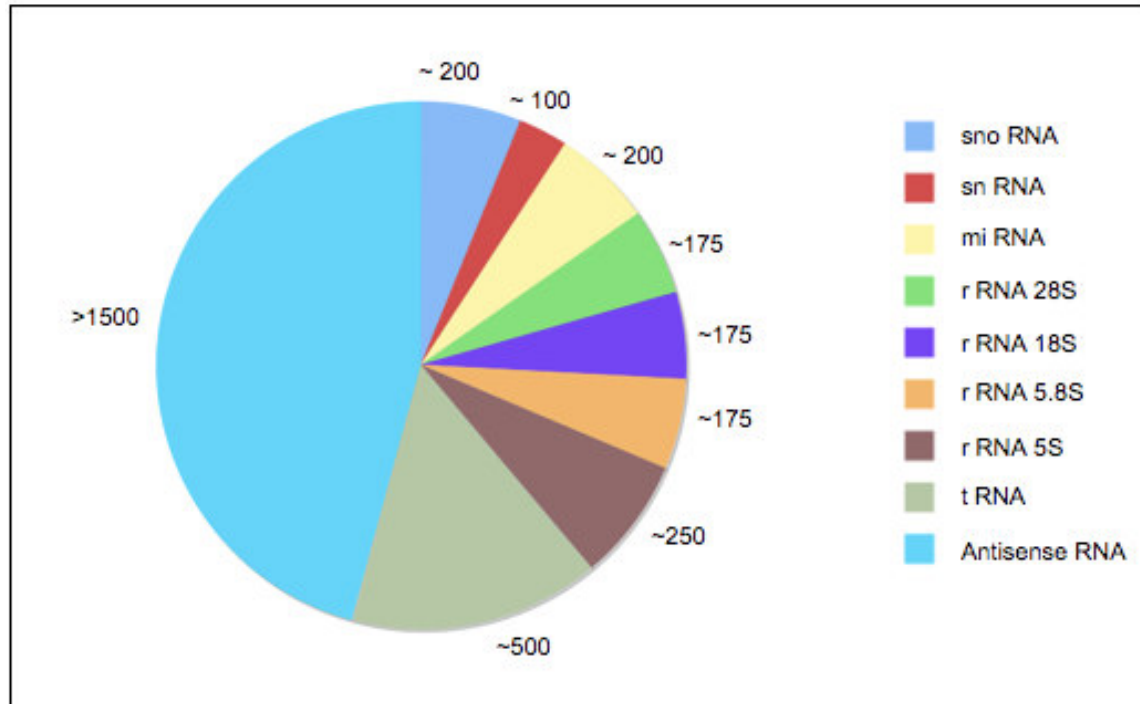


Figure 7. Non-coding RNA transcripts.

Micro RNAs and antisense RNAs are underestimated. Other non-coding RNAs are not present. (Taken from Human Molecular Genetics 3/e).

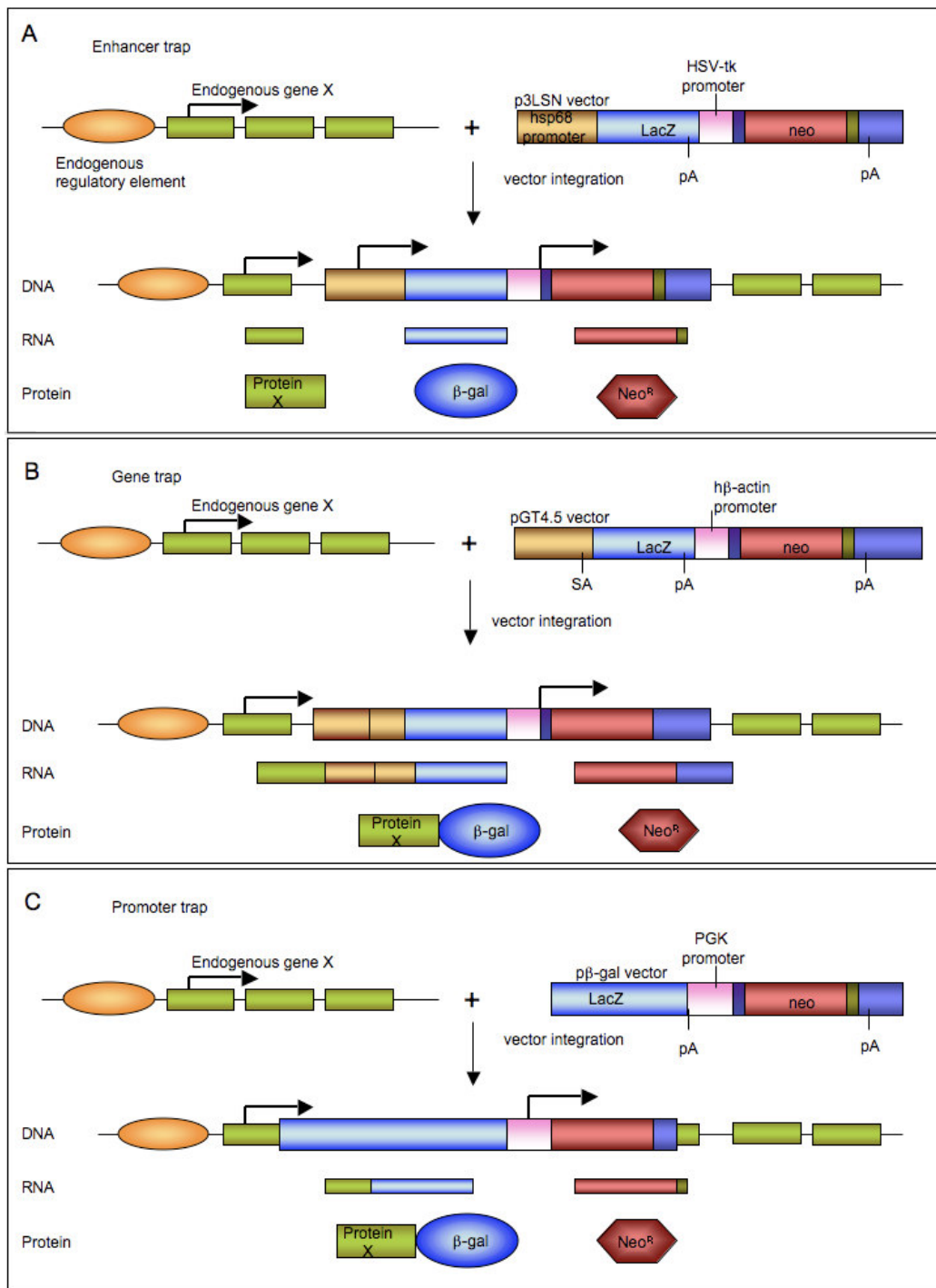


Figure 8. The basic trap vectors.

Enhancer-, gene-, and promoter-trap vectors contain a LacZ reporter gene and a neomycin resistance gene (neo) that is driven by an autonomous promoter, are shown trapping an endogenous gene “X”. Integration of the trap vectors into the ES cell genome will lead to neomycin selection whether the insertion has occurred inter or intragenically. A) The enhancer trap vector contains a truncated heat-shock inducible

minimum(hsp68)promoter upstream of LacZ. The insertion of the vector close to the enhancer of the gene X will lead to the transcription and translation of the LacZ reporter when the enhancer of gene X is activated. B) The pGT4.5gene-trap vector contains a splice acceptor (SA) site immediately upstream of a promoter less LacZ gene. Its integration in an intron leads to a fusion transcript generated from the upstream exon of gene X and LacZ upon transcriptional activation of gene X. C) The promoter-trap vector needs to be inserted into the coding sequence of gene X to activate transcription of the LacZ. On activation of gene X, a fusion transcript and protein between the upstream gene X sequence and LacZ will be generated.

MATERIALS AND METHODS

RNA extraction from ES cells

RNA was extracted from ES cells using the Trizol reagent (Invitrogen). Cells were plated on 100 mm dishes. 48hrs later cells were washed with PBS. After washing the cells were resuspended in 1 ml Trizol reagent. The samples were maintained at room temperature for 3 minutes. 0.2ml of chloroform/ml Trio, were added, the samples were mixed gently, and they were maintained for 10 minutes at room temperature. The resulting solution was centrifuged at 14,000 rpm for 15 minutes at 4°C. The supernatant was discarded and the pellets were washed with 1 ml 75% ethanol. After washing the samples were air dried for about 10 minutes. The pellets were than resuspended in proper amount of DEPC treated water.

DNase digestion

The RNA samples were treated with 10 U DNase I (Ambion) per ml RNA sample at 37°C for 30 minutes. The digested samples were treated with a DNase. RNA was then checked on 0.8% agarose gel and quantified by measuring the absorbance at 260 nm. An absorbance of 1 unit at 260 nm corresponds to 40 µg of RNA per µl.

cDNA transcription

cDNA synthesis was performed using the SuperScript™ First-Strand Synthesis System for RT-PCR (Invitrogen). cDNA syntesis was performed using random hexamers as follows:

| | |
|-----------------|-------|
| RNA | 1µg |
| Random hexamers | 0,5µl |

| | |
|--------------------|---------------|
| 10mM dNTP mix | 1 μ l |
| DEPC-treated water | to 10 μ l |

Each sample was incubated at 65°C for 5 min and incubated on ice for 1 min.

Then 9 ml of the following reaction mixture were added to each sample:

| | |
|------------------------|-----------|
| 10X RT buffer | 2 μ l |
| 25mM MgCl ₂ | 4 μ l |
| 0.1M DTT | 2 μ l |
| RNase inhibitor | 1 μ l |

Samples were incubated at 25 °C for 2 min, and then 1 ml of Superscript™ II RT was added to each tube. Samples were then treated as follows:

42°C for 50 min

70°C for 15 min

Samples were then chilled on ice and treated with 1 ml of RNaseH at 37 °C for 20 min.

The cDNA synthesized in this manner was used as template in PCR experiments.

The PCR mixture was prepared as follows:

| | |
|-------------------------|-------------|
| Buffer10X | 5 μ l |
| MgCl ₂ 25 mM | 3 μ l |
| dNTPs 5 mM | 2 μ l |
| Primer Fw 10 μ M | 2 μ l |
| Primer Rv 10 μ M | 2 μ l |
| Takara la Taq (~ 5 U) | 0,4 μ l |
| RNA (50 ng – 100 ng) | |

Add sterile H₂O to a final volume of 50 μ l.

Program

1. 65 °C for 10 min
2. 95 °C for 1min
3. 58 °C for 1min
4. 72 °C for 1min

repeat for 35 cycles steps from 2 to 4.

5. 72 °C for 7min
6. 42 °C for 60min
7. 70 °C for 15min.

Analyse the PCR product by gel electrophoresys.

Real Time quantitative PCR

A 2x PCR supermix from Bio-Rad (iQTM SYBR[®] Green supermix) containing Taq DNA polymerase (iTaqTM polymerase), MgCl₂, dNTPs, SYBR[®] Green, and fluorescein was used. Primers were added to the reaction mix at a final concentration of 400 nM. 1 micro g of total RNA purified from ES cells and DNase digested was reverse transcribed as previously described. The cDNA was added at a dilution of 1:3. For each sample, three distinct amplifications were carried out in parallel. The real-time quantitative RT-PCR was performed using an iCycler iQ system (BioRad). Cycling conditions were: 3 min at 95° followed by 40 cycles of 10 sec at 95°, 30 sec at 60° and 45 sec at 72°. The fluorescence data used for quantization were collected at the end of each 72°C step, and the treshhold cycle (ct) was automatically determined using the accompanying iCycler iQ software by calculating the second derivative of each trace and looking for the point at maximum curvature. GAPDH was used as a reference gene.

Agarose gel electrophoresis

Agarose gels (1 % w/v in TAE; 40 mM Tris-acetate pH 7.5, 2 mM EDTA) were prepared and supplemented with ethidium bromide (ca. 1 µg/ml). The percentage of agarose in gels was determined depending on the size of the DNA fragments to be resolved. Gels were generally run at 120 V in 1x TAE buffer, and DNA was visualised on a UV transilluminator.

DNA sequence analysis

For DNA sequence analysis, 100 ng of the PCR products were air dried and sent for sequencing.

Isolation of DNA from agarose gels

Following agarose gel electrophoresis, DNA gel slices were excised under UV light. DNA was extracted from these gel slices using Qiaquick columns (Qiagen) following the gel extraction protocol supplied by the manufacturer. Purified DNA was eluted from the columns using 30-50 µl of deionized water.

Cloning of the PCR products

2ml of the PCR product was added at the following mixture:

1ml of TOPO salt solution

0.25 ml TOPO vector

2.75 ml sterile water

The mixture was incubated for 30 minutes at room temperature. During this step an aliquote of competent XL1 blue bacteria (strain suitable for blue/white screening) was thawed on ice. The TOPO ligation was chilled on ice for 5 minutes, then 4 µl of this

ligation was transformed by heat shock into competent cells. The transformation was spread on to LB + amp plate.

In vitro transcription Materials

- DNA template (purified PCR product or linearized plasmid)
- DIG RNA Labelling Mix (Roche)
- Ribonuclease inhibitor (Fermentas)
- RNA polymerases
 - Sp6 (New England Biolabs)
 - T7 (New England Biolabs)
 - T3 (Stratagene)
- 10x transcription buffer (supplied with the RNA polymerase enzyme)
- DNase I, RNase-free (Roche)
- 1.5ml safe-lock, RNase-, DNase-, ATPase-free microtubes (Eppendorf Biopur)
- 0.3M MgCl₂ (it is practical to prepare a 3M stock solution and dilute 1:10 before using)
- 4M NH₄Ac, autoclaved (keep at -20°C)
- 100% Ethanol (keep at -20°C)
- 70% Ethanol (keep at -20°C)
- DEPC-treated water

In vitro transcription

The following mixture was prepared per reaction:

20 µl of DEPC treated water

2 µl of 10x transcription buffer

2 µl of 10x DIG RNA labelling mix

1 µl RNase inhibitor

1 µl RNA polymerase (T7 or SP6)

1µg of linearized DNA plasmid.

Samples were incubated at 37°C for 2.5 hrs. In the meantime, the Stop Solution was prepared as follows:

16.4 µl of DEPC-treated water

1.6 µl of MgCl₂ (0.3 M)

2.0 µl of DNase I (10 U/µl)

1µl Stop Solution was added per reaction. This aids to stop the IVT and remove the DNA template.

Samples were incubated at 37°C for 15 min.

Precipitation of RNA

To precipitate the RNA 72 µl ice-cold NH₄Ac (4 M, autoclaved) and 470 µl ice-cold 100% EtOH were added.

Samples were placed at –80°C for 20 min, then centrifuged at max. speed in a table-top microcentrifuge (e.g. 13,000 rpm) at 4°C for 20 min. Supernatant was removed carefully making sure not to disturb the pellet. The pellet was washed with 640 µl 70% EtOH, then centrifuged at maximum speed at 4°C for 20 min. Then the supernatant was removed carefully. The pellet was dried in a vacuum centrifuge for approx. 6 min to eliminate ethanol residue which may interfere in later reactions.

The pellet was re-suspended with 22 µl DEPC- H₂O shaking (in horizontal shaker) at 1150 rpm for 15 min (room temperature).

Quality control and quantification

Samples were checked on a 1% agarose gel. 1 µl riboprobe was added at 4.0 µl DEPC-H₂O and denatured for 5 min at 70°C.

The samples were chilled on ice for 3 min, and 1 µl 6x loading buffer was added to each sample. Then samples were loaded on a 1% agarose gel and an electrophoresis was run. 1µl of probe was diluted 1:100 (in Tris-buffer) and its concentration was determined using a spectrophotometer ($A_{260/280}$).

$[\text{ng}/\mu\text{l}] = (A_{260}) \times (\text{dilution factor}) \times (40)$. Concentrations between 500 – 850ng/µl are usual.

Prior to using for *in-situ* hybridization, riboprobes are further diluted to a concentration of 30 ng/µl in hybridization-mix and stored at –20°C for a maximum of 2 months.

Digoxigenin in situ hybridization

Preparation of tissues:

Wash the embryos in cold PBS, then transfer to fresh 4% paraformaldehyde/PBS at 4°C over night. Then transfer to 30% sucrose /PBS until they reach the bottom of the solution. Transfer embryos to a mixture of 30% sucrose/PBS and OCT at a 1:1 ratio agitating gently for 2 hrs at RT. Trasfer embryos to chilled OCT and freeze on dry ice. Store the sample at -80°C.

Preparation of sections:

10µm cryosections were collected on superfrost plus slides, the section were dried overnight at RT, and used the next day.

Pretreatment and hybridization:

Fix in fresh 4% PFA at RT for 15 min. Wash in PBTfor 5 min at RT. Bleach with 6% H₂O₂/PBT for 5 min at RT. Wash 3x PBT at RT for 5 min. Add 1 µg/ml proteinase K/PBT for 15 min at RT, then wash the embryos sections with fresh 2 mg/ml lysine /PBT for 10

min at RT. Wash 3x PBT. Prehybridize at 65°C for 1 hr. Hyb at 65°C overnight in closed containers.

Replace the slides in prewarmed (at 65°C) the sol1 (phormammehyde 50%/SSC/ SDS) for 15 min at 65°C. Repeat the wash in sol3 for 3 times. Wash in TBST for 10 min at RT.

Block in 10% sheep serum/MABT for 1hr at RT. Incubate with anti-digoxigenin antibody (1:2000) in 1% sheep serum/MABT at 4°C overnight.

The morning after wash for 4 times in TBST for 15 min at RT. Then wash with NTMT at RT for 10 min. Incubate with NBT/X-phos in NTMT in the dark looking at the signal. Wash twice with NTMT at RT for 10 min. Then wash in PBS1X at RT for 5 min for two times. Put the samples in 4% PFA for 30 min a RT. Finally wash for 5 min in PBS1X. Repeat the washing. Section were mounted using Glycerol 70%/PBS and examined with an Axioplan microscope (Zeiss) equipped with an AxioCam CCD camera and Axiovision digital imaging software (Zeiss).

Obtaining zebrafish embryos

Natural cross fish

The afternoon before the cross set up fishes. Place a smaller plastic container with a mesh bottom into a larger container. Add fish water to cover for some inches of water the mesh. Transfer a pair of fishes into the container. After the light comes on wait a bit and then cross the fishes. The onset of light is a major stimulus for zebrafish to breed. Collect the eggs from the bottom container. After the egg collection, separate the embryos and wash them in a Petri dish. The fish may lay a larger number of eggs comprised between 30-50 eggs.

Removing the chorion

Transfer the embryos using a plastic pipette. The embryos are still in their chorions. The chorions removal occurs pretreating the embryos with 0.5 ml of a 10 mg/ml pronase in water for 1min in a Petri dish. It's important watching the embryos. As soon as their chorions start to bubble change Petri dish. Stir gently. Thus, when the first 3-5 embryos are released from their chorion decant the content into a 500 ml backer filled with fish water. Now the embryos are extremely fragile. Repeat the washing for two times.

Once chorion is removed put the embryos using a plastic pipette in agar coated Petri dishes cause the embryo have to stay away from water and plastic surface cause they could explode.

Preparation of the injection solution

Purify PCR fragments using the Qiagen PCR purification kit. Run a 1% agarose gel and quantify the PCR products. The injection solution have to contain 1/10vol of phenol red DNA fragments at a concentration of 50 ng/ μ l reaching a range of 5 to 1molar ratio with the HSP LacZ fragment.

Centrifuge for 2 min at maximum speed the injection solution in filter column contained into a sterile eppendorf tube to remove particle debris that could block injection needles.

Microinjection of early embryos (1-2 cell stage)

Use a needle puller to prepare glass microcapillary needles.

Add 1 μ l of injection solution using a pipette in microfilament containing needles.

Set presure conditions for injection: Pressure 10-200 psi, time: 0.3 ms.

Place embryos in agarose plate in 10 Hank's solution under the stereomicroscope.

Place needles into the embryo without shaking. Inject a drop with a diameter approximately $1/10^{\text{th}}$ of the diameter of the animal pole of the embryo. For injection use a pedal (Narishige Harvard Scientific).

Put the embryos at 28°C to let them to develop. After 4 hrs look at the embryos and remove the ones that develop abnormally.

LacZ staining

After 24 hrs place the embryos into a 24 well plate in Hank's buffer. Then replace the Hank's buffer with a BT-Fix and let the embryos stay at 4°C in BT-Fix for 4 hrs. Then wash 3-5 times. Wash for 5 min with the Staining Buffer. Use 1 ml Staining Buffer + 5µl 8% X-Gal in DMSO for the staining. Wash the embryos for 3 times with PBS/ 0.02% NP40. Fix over night with BT-Fix at 4°C. Wash them again in PBS/ 0.02% NP40. Draw the expression maps.

RESULTS

Background

One of the aims of this thesis was the identification of novel genes using gene trapping as a novel approach to re-annotate the mouse genome.

To start identifying novel genes 249,827 traps were collected from several public and private gene trapping projects found within the GSS section of GenBank. Among these sequence tags, 95.2%, defined as “mRNA traps”, were obtained by 5’- or 3’- RACE-PCR of the fusion transcript between the reporter gene and the endogenous gene while remaining sequences, named “genomic traps”, revealed the exact genomic insertion site because the sequence was obtained by inverse-PCR.

These sequence tags were mapped to the genome using a stringent pipeline. This analysis showed that while 65% of them found a clear localization in the genome, 26% presented an ambiguous mapping that maybe due to the poor quality of the sequences, and the remaining 9% did not find any match in the genome. In fact in most of the cases (43%) traps had a sequence length shorter than 50 nucleotides, making it difficult to assign them to an exact genome location. Other reasons that explain the lack of mapping for the remaining 5% of traps are the presence of spurious sequences in the data set as well as genome coverage issues.

Traps were subsequently assembled on the basis sequence overlap: if two traps overlapped on the same strand of a chromosome by at least one base pair, they were put together in a cluster (named hereon “trapclusters”). About 12,509 traps indicated spliced

transcripts and were used to verify the presence of canonical splice site junctions in order to have some indication of the existence of a putative gene.

In order to investigate if these sequence tags are able to detect novel genes they were compared with available collections of transcribed sequences, such as FANTOM3, based on full-length cDNAs (Carnici et al., 2005), and Unigene, which is based on clusters of EST sequences (Schuler et al., 1997). Interestingly trapclusters presented the highest proportion (40%) of unique sequences among the three data sets. This result suggests that the ES cell transcriptome might contain additional information, e.g. molecular features specific to their totipotency state, which are quite different from those obtained by FANTOM3 and Unigene in different tissues and cells. Comparing our dataset to the RefSeq database we observed that 44% of the trapclusters overlapped with a known RefSeq gene. Moreover, among those which do not overlap RefSeq, 9% overlapped with genes predicted by Ensembl but not found in RefSeq, a further 7% with cDNAs identified by Fantom3 and not present in Ensembl, and finally a further 2% with EST clusters contained in Unigene but nor present in the above datasets. Overall 38% of the trapclusters identified indicate novel putative features of the transcriptome that have never been annotated before.

Novel exons within known genes

Having mapped our dataset to the genome, and having identified a subset which did not overlap known gene databases we investigated whether this dataset was in fact adding novel putative exons to currently annotated RefSeq genes, which would better refine their currently annotated structure, taking into account that RefSeq contains curated gene structures which have been verified experimentally. We therefore investigated trapcluster

sequences showing a partial overlap with current RefSeq gene structures that could indicate novel potential exons. The analysis identified internal exons, as well as 5' and 3' exons on 830 RefSeq genes (Fig. 9). In order to verify the existence of these candidate exons, we chose 40 of these and we performed RT-PCR experiments using as template ES cell RNA to verify the expression of these exon.

In these experiments we decided to project a primer on a candidate exon and a primer on the closest exon belonging to the annotated gene. We confirmed the existence of the predicted exons in 40% of the cases. Furthermore we verified if these novel exons were specifically expressed in ES cells or whether they could constitute alternatively spliced exons which would occur only in specific tissues. For this reason we performed the same RT-PCR analysis using total RNA extracts from several mouse tissues, i.e. adult brain, eye, heart, and whole embryo at 14.5 days of development (E14.5). This additional verification confirmed that indeed these exons presented complex patterns of splicing, but also showed that some exons which are trapped in gene trapping experiments are not necessarily expressed at detectable levels in ES cells. This additional verification confirmed as expressed a further 30% of the exons predicted. Thus the compound result of these verifications, taking all tested tissues into account, yields an overall success rate of 70% (Table1). Table 1 shows all the exons which were tested by RT-PCR across different RNA samples. This test allowed us to group exons in different categories depending on their expression pattern. Exons which were found to be expressed only in ES cells were named “ES-only”, those which were expressed in all tissues tested were named “ubiquitous”, those which showed a complex on-off pattern of expression and different amplification products of several lengths depending on the RNA used were

named “complex”. Finally the category named “ES-absent” comprises 12 exons which could not be detected in ES cells. Six of these exons were trapped using a polyA-type vector which is able to trap genes even if they are not transcribed in ES cells, and the other six were trapped by an SAbgeo-type vector, thus they are probably expressed at very low levels in ES cells and are up-regulated upon differentiation. The last group was named “absent” as it contains exons which could not be verified in any of the RNAs tested. We cannot exclude that these exons could be real and could be expressed in tissues/stages which we did not test. These data taken together demonstrates that gene trapping is able to capture both expressed and non-expressed genes, depending on the type of vector used. Figure 10 shows some examples of known genes to which our analysis added novel exons. For example, an alternative 5’UTR exon was added to the known *Ncapg2* gene (Fig. 10 A), which shows a complex expression pattern, since it appears to yield several splicing variants depending on the RNA sample used. Similarly trapcluster TCL606 (Fig. 10 B) indicates a new 5’UTR exon within the *Niban* gene, which is expressed in a stage specific manner, as a clear band can be observed in ES cells and in whole embryo, while in other tissues it is not possible to detect any signal.

In the case of trapcluster TCL195 (Fig. 10 C), we verified the addition of a novel internal exon between exon 3 and exon 4 of the known gene *Nol1* and it was found to be expressed in all samples tested, giving always the same product length. This suggests that there is an alternative transcript of *Nol1* containing this novel exon that had not been observed before. On the contrary, the addition of a new exon to the *Inpp5d* gene occurs only in ES cells among the samples tested, suggesting that this alternative transcript could, perhaps, have a specific role in ES cells.

A further group of exons tested fell in the 3'UTR of known genes (Fig. 10 E, F). The TCL10445 cluster (Fig. 10 E) seems to add a 3'UTR exon to the *Rheb11* gene. However, when sequencing this alternative product we realized that the resulting transcript, which does include this novel exon, skips the last two exons of the gene in all RNA samples, except in whole embryo. Finally we also show the addition of a 3'UTR exon to the *Bcl7c* gene. This exon is expressed only in ES cells and is located quite far, at a distance of 30 kb from the last known exon of the gene.

Identification of Novel Transcripts

We observed that a large number of trapclusters (~66%) did not overlap with other clusters or known genes, so it was difficult to hypothesize the start and the end of a putative gene embedding the trapcluster. Thus we used CpG islands and transcription start sites (TSS) predicted by Eponine (Down et al., 2002) to define potential gene boundaries around trapclusters, to reduce this large data set into a lower potential number of novel genes. In this way, it was possible to group adjacent (but not overlapping) trapclusters into a set of about 8,420 novel transcripts classified into 1,997 “novel genes” (found in regions between CpG islands without any annotation) and 6,423 “novel transcripts” located within known transcriptional forests (Fig. 11). About 1,333 are “nested”, that means that these genes are in the same direction of a known transcript within the locus but are completely contained within its introns, while 792 were considered putative “antisense transcripts” because they have an orientation that is opposite with respect to a known transcript.

We choose 80 random transcripts (1%) from this reduced dataset and we proceeded to verify their existence as well as that of all of the exons contained within them by RT-

PCR experiments. We found that 71% of these genes (57/80) are expressed in ES cells (Table 2), and we also confirmed that 50% of their exons are expressed in ES cells. As a negative control we performed 10 RT-PCRs using 20 existing trap primers assorted randomly while as a positive control we performed a similar analysis using primers for trap TCLG4070. While the positive control was confirmed, all other primers gave negative results.

Fig. 12 A shows an example of the results obtained, indicating the TCLG1417 transcript, which lacks an ORF, thus probably a non-coding gene. Interestingly, this novel transcript was identified to be in opposite orientation and partial overlap with respect to the known gene *Trpm3*, indicating a potential regulatory role. This new non-coding gene was predicted to have 10 exons and our RT-PCRs confirmed 7 out of 10 of these exons as truly expressed in ES cells.

Another predicted transcript which was found in reverse orientation with respect to a known gene is TCLG1647. In this case, the predicted gene is actually larger than the known gene (*Tcf15*) which is fully contained within one of its introns. This gene also appeared to contain 7 exons, but the PCR analysis showed that among the predicted exons two proximal exons form, in reality, one larger exon. Moreover, the sequence data brought about the addition of yet another exon that was not present in the trap collection as well as two separately expressed exons that could not be linked to this transcript.

TCLG400 (Fig. 12 C) is found in opposite orientation and partial overlap to the *Ngfr* gene. It is composed by 4 exons, which were all confirmed by RT-PCR. TCLG1753 (Fig. 12 D) is a single gene, antisense to the *Prkci* gene, in which we confirmed 3 out of 5 predicted exons. We also confirmed all five exons predicted for the TCLG2423, three of

which show a partial overlap with the *1110032016Rik* gene. Finally Fig. 12 E shows the TCLG4470 gene, which contains 3 exons and is found in opposite direction and fully contained (i.e. nested) within the intron of the *Oprd1* gene.

Expression profiling of a non-coding transcript.

In order to verify further the expression of non-coding transcripts within our data set, we performed an in situ hybridization of one of the transcripts verified by RT-PCR, TCLG1417, the non-coding gene antisense of *Trpm3* described above (Fig. 12 A), on a mouse embryo at 14.5 days of development. This developmental stage is very representative because it represents an interesting temporary window in which a large number of genes are expressed. The results from the *in situ* hybridization indicated a very specific pattern of expression in the inner ear (chochlea and vestibules), in the choroid plexus and in the eye (Fig. 13). The same expression pattern was obtained in E12.5 embryos and in P0 mouse (data not shown).

These data led us to hypothesize that this novel transcript could act as antisense regulating the mRNA stability of the *Trpm3* gene. *Trpm3* is a poorly understood member of the large family of transient receptor potential (TRP) ion channels. In literature five splice variants have been reported. In situ hybridization experiments conducted on this gene showed that *Trpm3* is expressed in several regions of the mouse brain such as the dentate gyrus, the intermediate lateral septal nuclei, the indusium griseum, and the tenia tencta (Oberwinkler et al., 2005) and northern blot analysis confirmed expression also in the eye. Interestingly, strongest expression was observed in epithelial cells of the choroids plexus. Further experiments will be focused on understanding the function of

this novel gene, its possible role in the auditory pathways as well as its potential interactions with the *Trpm3* gene.

Trapping of genes correlates with their expression levels

Although the majority of genes is trapped only once or very few times, a small subset of genes is trapped hundreds of times. Therefore we decided to investigate whether this subset of genes also displays significantly higher levels of expression in ES cells. One factor that could theoretically influence the rate of insertion and thus that of trapping events, is the accessibility of the chromatin of the genomic region. For this reason, genes which are involved in transcriptional pathways, for example, could be reached more easily by DNA vectors used for mutagenesis. These regions are considered “gene trapping hot spots”. These regions have been observed before but have never been investigated in further detail (Hansen et al., 2003). The distribution of these regions across the genome appears to be random and uniform. Another factor that could influence the rate of gene trapping is the overall size of the gene locus because the more space is available for insertion of the vector, the more likely the event is to occur. Thus we ranked genes and trapclusters according to the number of trapping events and normalized the dataset for overall locus size in order to identify genes that are likely to be trapped at high rates due to their expression levels. In this manner we selected 383 genes which we defined as being “hypertrapped”. The first 50 genes are shown in Table 3. This number represents less than 5% of all the genes trapped but contains more than 20% of all gene traps sequenced.

To test if hypertrapped genes indicate genes that are highly expressed in ES cells, we performed a real-time PCR experiment. We chose 10 genes from the hypertrapped

gene list and, as a control, we also chose randomly 10 genes that were trapped once (the median rate of trapping). Moreover we compared the expression levels obtained for these genes in ES cells to the *Oct4* gene, a well known marker of these cells (Niwa et al., 2000). The results show that 80% of the hypertrapped genes we tested were expressed at levels significantly higher than the control set and also that these levels of expression were comparable to those of *Oct4* (Fig. 14). These data confirmed that this set of genes is significantly expressed in ES cells at levels similar to those of a gene which is known to play an important role in these cells. A comparison with other previously published datasets of expression profiling in ES cells revealed remarkably low overlap. Among the genes tested, only one, *Scpep1*, is present in two of three previously published data sets. This is a serine carboxylpeptidase which takes part in the activation of other proteins after a proteolytic cut. Immunohistochemical studies on this protein showed that *Scpep1*, while in the embryo, is expressed in the heart, vascular apparatus and the aortic skeletal muscle, while in the adult it is expressed in endothelial cells (Lee et al., 2006).

Another hypertrapped gene is *Mshi2h* (Musashi homolog 2). It has been shown recently that this gene is involved in the maintenance of ES cell identity (Siddal et al., 2006) although it is not found to be significantly expressed in expression profiling studies published until now. Other hypertrapped genes all appear to be involved in early development. They include: *Erdr1* (erythroid differentiation regulator 1), which is known to be highly expressed in the early phases of erythroid lineage development and in cephalic mesenchyme development, just like *klf9* (Krueppel like factor 9) (Martin et al., 2001); *Gabarapl2* (GABA receptor associated protein like 2) is highly expressed in the early developmental stages of the neural tube and the notochord (Liang et al., 2004); *Rbpms*

(RNA-binding protein with multiple splicing) is involved in heart development (Gerber et al., 1999) and *Cfdp1* (Craniofacial development protein 1) in craniofacial development (Mukhopadhyay et al., 2006). Other hypertrapped genes are involved in chromatin remodeling, such as *Cbx5* (Chromobox protein homolog 5) (Yamaguchi et al., 1998), in protein folding, such as *Pfdn1* (prefoldin 1) (Zako et al., 2005) or in ubiquitination pathways, such as *Ube2r2* (ubiquitin-conjugating enzyme E2R2) (Semplici et al., 2002).

These results suggest that hypertrapped genes constitute a novel set of genes that are expressed at significant levels in ES cells and might be relevant to clarify further mechanisms that characterize these cells.

Verification of shuffled conserved elements in the vertebrate lineage

In the second part of this work we investigated *in vivo* the function of shuffled conserved non-coding elements. These elements, conserved across the vertebrates, were identified using a combination of different tools. Firstly, orthologous loci from four mammalian genomes were used to identify “rCNEs”, i.e. regionally-conserved elements. Subsequently, these rCNEs were compared with orthologous loci in fishes to investigate if the conservation was also extended in these organisms. In this way, we identified shuffled conserved elements (SCEs), i.e. regions of the mouse genome conserved in the *Takifugu rubripes* orthologous locus with 40bp length and 60% conservation. Thus 21,427 non-redundant, non-genic, shuffled conserved elements were found in 30% of the genes analyzed (2,911). Only 28% maintained the same orientation and the same position with respect to the gene and were name “collinear”, while the remaining SCEs were shuffled, i.e. have either changed orientation or position or both between the mouse and fugu genome during the evolutionary time separating these two organisms (Fig. 15). We further proved that the extent of shuffling observed was not due to an assembly artifact by verifying the collinearity independently in two fish genomes (*Fugu* and *Tetraodon*). Moreover we showed that conserved elements are significantly more often collinear in the 500bp window adjacent to the TSS of the gene as compared to any other analyzed region, probably owing to elements which are position and orientation constrained in the core promoter region.

Verifying SCE function

To investigate a putative function for the SCEs identified, we performed an overlap analysis with 98 known mouse enhancer elements annotated in Genbank. The overlap of SCEs was compared with the overlap of two other datasets of conserved non-coding sequences which show conservation in fishes. Interestingly, we showed that the SCEs dataset overlapped with 18 known enhancers while the CNE and UCE datasets overlapped with only 1 and 2 known enhancers respectively.

To corroborate these findings and validate the enhancer activity of these SCEs, we screened these elements in zebrafish embryos. Thus, DNA fragments amplified from the fugu genome were purified and then co-injected using a construct containing the *shh* (sonic hedgehog) promoter and *LacZ* as a reporter gene. The co-injection was performed into zebrafish embryos at the early stages of development (1-2 cell stage) and after 24 hrs of development *LacZ* expression was observed. We counted about 60 embryos for each injected DNA fragment. We tested 27 fragments, 4 of which overlapping with known mouse enhancers that have never been tested in fish before. The remaining 23 did not overlap to any known feature.

We also injected, as a control, 12 non-coding, non-repeated and non-conserved fragments, 9 of them were from the same genes from which SCEs have been chosen while the remaining 3 were from random genes. As previously reported, this type of analysis is characterized by significant mosaicism of the expression of the transgene (Westerfield et al., 1992). To obtain an expression profile of the enhancer activity, we counted the number of cells stained for X-gal and we annotated the position of the expressing cells from a large number of embryos on expression maps. Expression maps represent a composite overview of the *LacZ* positive cells of all embryos tested. We

found that, when compared to the embryos injected with only the hsp lacZ construct, 22 out of the 27 fragments tested showed a clear enhancer activity, 3 fragments out of the 4 known mouse enhancers conserved in fish also act as enhancers in fish. Interestingly, we observed that the enhancer effect for each of the fragments tested was tissue specific, not generalized.

We also examined expression data from the Zebrafish Information Network (ZFIN) to compare the results obtained with the expression pattern of the genes neighboring the elements tested. Interestingly, several SCEs belonging to a single gene locus showed similar tissue specific activity. For example, we tested 4 different fragments belonging to the *ets1* locus. For all these fragments we observed a high specificity for blood precursors (see Fig. 16, SCE 1646). This finding corresponds with previously reported data which show that *ets1* is expressed in the venous and arterial system. Moreover, both fragments tested from the *zfmpm2* locus (*fog2*, Walton et al., 2006) showed specific enhancer activity in the CNS, in line with the expression of both *fog2* paralogs that is brain specific (Walton et al., 2006). The fragment tested for the *mab-21-like* genes had specific enhancer activity in the CNS and in the eye (SCE 4939). This expression mirrors the pattern previously observed in the brain, eye and neurons (Kudoh et al., 2001; Kudoh et al., 2001b). SCEs from *pax6* and *hmx3* genes showed enhancement specific to the CNS, which also corresponds to previously reported expression patterns for these genes (Sprague et al., 2003). An SCE located in the *jag1b* locus showed specific expression in the CNS and in the eye. This result is only partially in line with the reported expression of this gene, which is reported to be expressed in the rostral end of the pronephric duct, nephron primordia, and in the region extending from the optic vesicle to

the eye (Zecchin et al., 2005). Moreover, we identified novel enhancer activity for the SCEs neighbouring *lmx1b1*, which showed CNS specificity, and SCEs found within *mapkap1*, *tmeff2* and *3110004L20Rik* (producing integral membrane protein) and *elmol* (associated with cytoskeleton), which showed strong generalized or tissue specific activity. For these genes there was no comparable expression data. In contrast only 2 of 12 (about 17%) of the control elements showed significant enhancement of *LacZ* activity (Table 4).

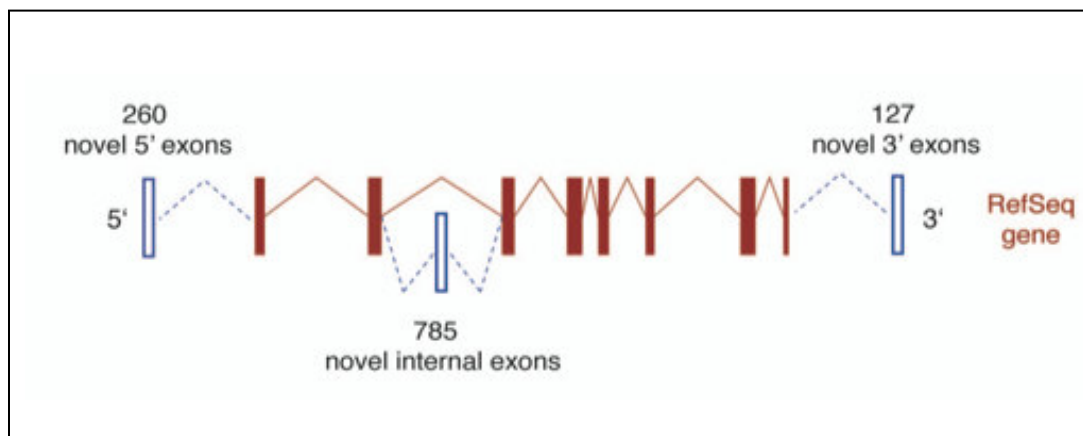


Figure 9. Prediction of novel exons (1172) identified on RefSeq genes.

The hypothetical exons are primarily external exons (785), as well as 5' exons (260) and 3' exons (127).

| | 5' exons | Internal exons | 3' exons |
|------------|---------------------------------------|--|--|
| ES-only | - | Inpp5d | NM_013842 |
| ES-absent | Inpp4a, Dpm3, Itsn1, 8430423A01Rik | Abcc1, Eng, Rnf111, Pip5k1a, 4931406I20Rik, Lasp1, Eif2ak3 | 9630015D15Rik |
| Ubiquitous | Nlgn3, Luzp5 | - | Rhebl1, D630023F18Rik, Srgap2, Armcx1 |
| Complex | Niban | 1110034C04Rik, Nol5, Anp32b, Slc6a6, Adck5, A930010I20Rik | Bcl7c |
| Absent | Rps21, Ssr2, D14Ert668e | Dnmbp, Adam23, Prkar2a, Dlgh3, Sec14l1, Aspscr1 | Srl, Tusc3, Tspan14 |

Table 1. Verification of 40 novel exons tested by RT-PCR using RNA extract from ES cells, whole E14.5 embryo, heart, brain and eye.

The table indicates exons which were found only in ES cell RNA as “ES only”, those that were absent in ES cell RNA but present in all other tissue as “ES-absent”, those that were detected in all RNAs tested as “ubiquitous”, those that showed complex on / off patterns and different products in the RNAs tested as “complex”, and those that could not be detected in any of the RNAs tested as “absent”.



Figure 10. Discovery of novel exons on known RefSeq genes.

Figure 10. Discovery of novel exons on known RefSeq genes.

The figure shows six samples of RefSeq genes to which were added novel exons using gene trap data, as well as images of the RT-PCR results in several tissue RNA and in ES cells RNA.

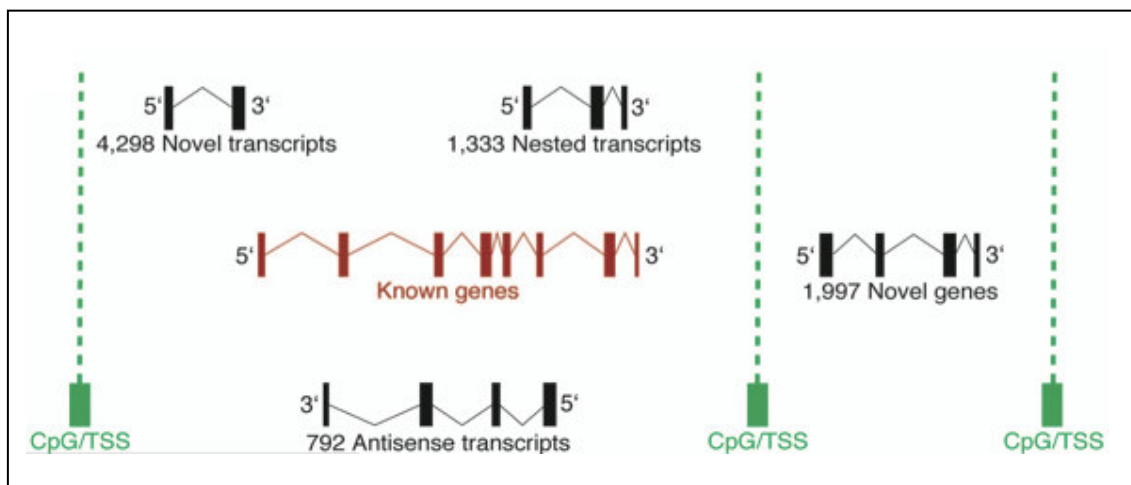


Figure 11. Prediction of 1,997 novel genes and 6,423 novel transcripts found within known gene loci.

1,333 novel genes are nested while 792 putative antisense.

| | Confirmed | Not Confirmed |
|--------------------------------|--|---|
| Nested <i>TCLG(gene)</i> | TCLG4845(2310001H13Rik), TCLG4470(Oprd1), TCLG4400(Akap2), TCLG4020(Kng2), TCLG3643(Spred2) | TCLG4185(Capn1) |
| Antisense <i>TCLG(gene)</i> | TCLG1647(Tcf15), TCLG400(Ngfr), TCLG3471(Slc25a5), TCLG1753(Prkci), TCLG947(Myo10), TCLG330(Myo1g), TCLG2538(1700016D06Rik), TCLG2221(Bcl7b), TCLG1581(Slc27a4, 2900073H19Rik), TCLG869(Slc1a3), TCLG486(Myo15, Drg2, 4933439F18Rik), TCLG2810(A230098A12Rik), TCLG2486(Alpk3, NP_001004184.1), TCLG2005(D030015G18Rik), TCLG1928(Arpm2), TCLG1590(NM_007494) | TCLG81(Gsta3), TCLG2356(Ddx47), TCLG2266(Spr), TCLG1688(Pag1), TCLG1004(Ankrd33, Acvrl1) |
| Novel <i>TCLG(Chr:Mb)</i> | TCLG2660(Chr8:88.23), TCLG2423(Chr7:120.93), TCLG2034(Chr4:147.31), TCLG724(Chr13:110.15), TCLG2616(Chr8:121.21), TCLG2519(Chr7:121.53), TCLG2033(Chr4:147.22), TCLG1131(Chr17:45.44), TCLG757(Chr13:90.82), TCLG467(Chr11:25.95), TCLG2808(Chr9:72.74), TCLG2022(Chr4:140.16), TCLG1541(Chr2:152.95), TCLG1153(Chr17:77.79) | TCLG978(Chr15:84.69), TCLG455(Chr11:3.18), TCLG2847(Chr9:120.76), TCLG1777(Chr3:89.95), TCLG1520(Chr2:103.51), TCLG1450(Chr19:52.61), TCLG1259(Chr18:36.46), TCLG1205(Chr17:45.52), TCLG1113(Chr17:25.43) |

Table 2. Results of RT-PCR verifications on ES cell RNA of 50 novel transcripts predicted to exist on the basis of gene trap sequence tags.

The table separates genes that were confirmed from those that were not confirmed. Moreover it separates transcripts that were found nested within known genes, antisense on known genes, as well other strand alone transcripts shown as novel.

Figure 12. Discovery of novel genes based on trapclusters.

Schematic representation of six samples of novel multi exon genes predicted using gene trap data and available CpG islands and Eponine transcription start site annotation, verified by RT-PCR on ES cell RNA.

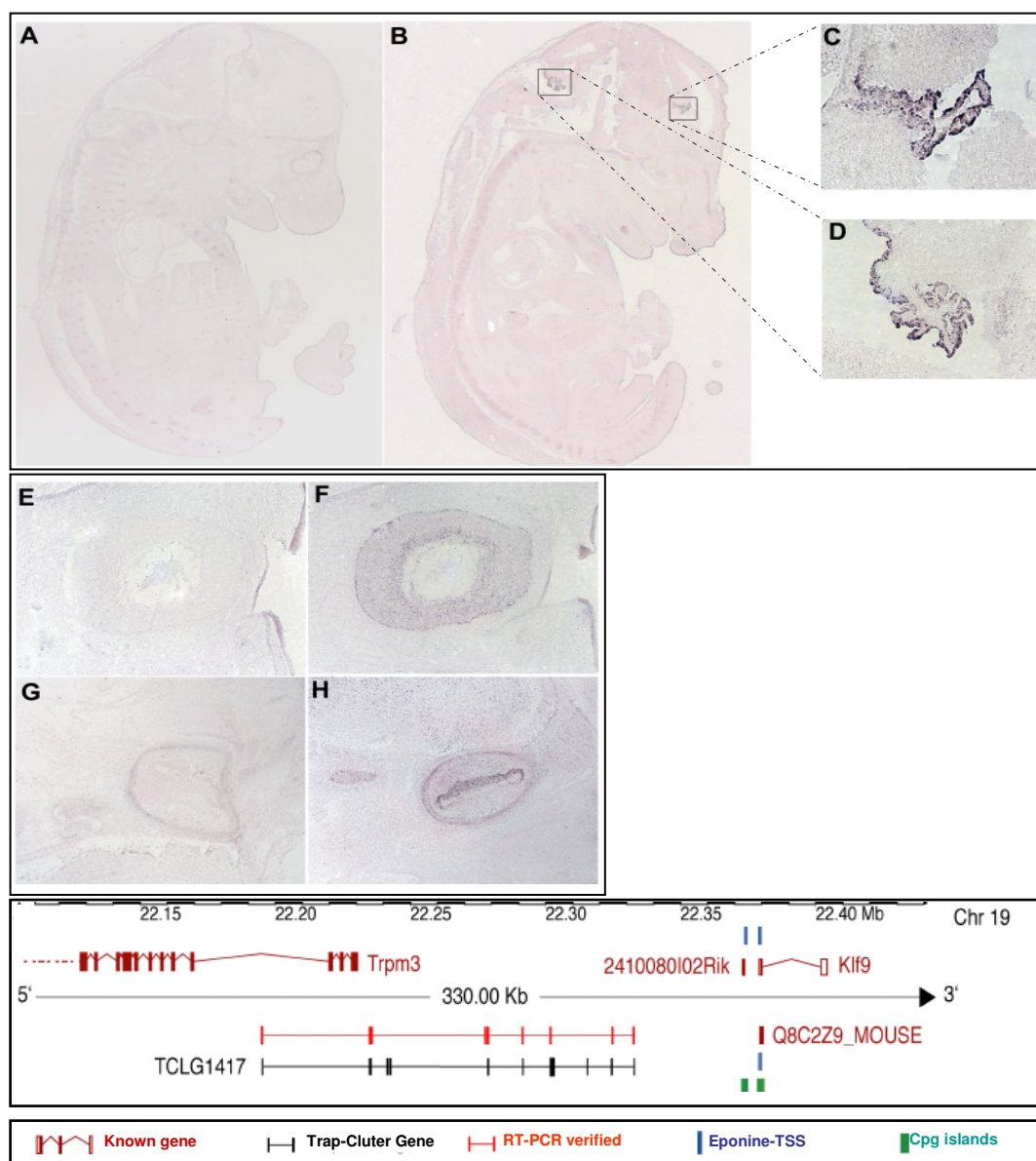


Figure 13. In situ hybridization of trapcluster gene TCLG1417 on E14.5 mouse.

This gene shows an highly specific signal within the III (C) and IV (D) ventricle of the choroids plexus, and in the eye (E, F) in the inner ear (cochlea G, H).

| Refseq_ID | Gene | Genomic localization | Description | Trap (n) |
|--------------|----------------|-------------------------------|---|----------|
| NM_133362 | Erd1 | ChrNT_051172:6351-7878 (-) | erythroid differentiation regulator 1 | 707 |
| NM_054043 | Msi2h | Chr11:88067453-88539178 (-) | Musashi homolog 2 (Drosophila) | 419 |
| NM_026027 | Pfdn1 | Chr18:36627482-36678298 (-) | prefoldin 1 | 323 |
| NM_011801 | Cfdp1 | Chr8:111066063-111151900 (-) | craniofacial development protein 1 | 270 |
| NM_008850 | Pitpna | Chr11:75313766-75354436 (+) | phosphatidylinositol transfer protein, alpha | 404 |
| NM_007626 | Cbx5 | Chr15:103258715-103303122 (-) | chromobox homolog 5 (Drosophila HP1a) | 446 |
| NP_062707 | Rbpms | Chr8:32588246-32735409 (-) | RNA binding protein gene with multiple splicing | 449 |
| NM_010638 | Klf9 | Chr19:22379148-22404833 (+) | Kruppel-like factor 9 | 162 |
| NM_029023 | Scpep1 | Chr11:88745108-88776520 (-) | serine carboxypeptidase 1 | 203 |
| NM_026275 | Ube2r2 | Chr4:41274873-41332222 (+) | ubiquitin-conjugating enzyme E2R 2 | 150 |
| NM_026693 | Gabaprl2 | Chr8:111238427-111253087 (+) | gamma-aminobutyric acid (GABA-A) receptor-associated protein-like 2 | 224 |
| NM_033327 | Zfp423 | Chr8:86945303-87244633 (-) | zinc finger protein 423 | 239 |
| NM_194059 | Nanos3 | Chr8:83436801-83439620 (-) | nanos homolog 3 (Drosophila) | 286 |
| NM_009980 | Ctbp2 | Chr7:127353899-127489680 (-) | C-terminal binding protein 2 | 189 |
| NM_013482 | Btk | ChrX:128087275-128128084 (-) | Bruton agammaglobulinemia tyrosine kinase | 431 |
| NM_018810 | Mkrl1 | Chr6:39533207-39555818 (-) | makorin, ring finger protein, 1 | 164 |
| NM_026391 | Ppp2r2d | Chr7:133232602-133269297 (+) | protein phosphatase 2, regulatory subunit B, delta isoform | 171 |
| NM_007632 | Ccnd3 | Chr17:45023592-45118144 (+) | cyclin D3 | 216 |
| NM_025927 | Mrpl45 | Chr11:97136944-97151008 (+) | mitochondrial ribosomal protein L45 | 145 |
| NM_008034 | Folr1 | Chr7:95964456-95976886 (-) | folate receptor 1 (adult) | 164 |
| NP_035727 | Tjp2 | Chr19:23332680-23378444 (-) | tight junction protein 2 | 266 |
| NM_145510 | Rabif | Chr1:134344924-134358149 (+) | RAB interacting factor | 135 |
| NM_198417 | C030039L03R | Chr7:23067546-23081244 (+) | RIKEN cDNA C030039L03 gene (C030039L03Rik), mRNA | 64 |
| NM_013625 | Pafah1b1 | Chr11:74399611-74450329 (-) | platelet-activating factor acetylhydrolase, isoform 1b, beta1 | 160 |
| NM_009456 | Ube2l3 | Chr16:15923030-15972516 (-) | ubiquitin-conjugating enzyme E2L 3 | 255 |
| NM_001003918 | Usp7 | Chr16:8364074-8464206 (-) | ubiquitin specific protease 7 | 299 |
| NM_023197 | 2310008H09R | Chr7:112719648-112732117 (-) | RIKEN cDNA 2310008H09 gene (2310008H09Rik), mRNA | 149 |
| NM_145823 | Pitpnc1 | Chr11:107032524-107158727 (-) | phosphatidylinositol transfer protein, cytoplasmic 1 | 226 |
| NM_026615 | 2900073H19R | Chr2:29759552-29777159 (+) | RIKEN cDNA 2900073H19 gene (2900073H19Rik), mRNA | 128 |
| NM_027230 | Prkcbp1 | Chr2:165242023-165353684 (-) | protein kinase C binding protein 1 | 351 |
| NM_016786 | Hip2 | Chr5:64339101-64400758 (+) | huntingtin interacting protein 2 | 121 |
| NM_183278 | 22000011I15Rik | Chr14:32488464-32491975 (-) | RIKEN cDNA 22000011I15 gene (22000011I15Rik), mRNA | 151 |
| NM_008692 | Nfyc | Chr4:119779884-119848112 (-) | nuclear transcription factor-Y gamma | 197 |
| NP_031523 | Atf1 | Chr15:100285518-100317872 (+) | activating transcription factor 1 | 185 |
| NM_026532 | Nutt2 | Chr8:105156480-105176250 (+) | nuclear transport factor 2 | 96 |
| NM_008942 | Npepps | Chr11:97028021-97101651 (-) | aminopeptidase puromycin sensitive | 158 |
| NM_009642 | Agtrap | Chr4:146569424-146580366 (-) | angiotensin II, type I receptor-associated protein | 102 |
| NM_007792 | Csrp2 | Chr10:110543028-110562471 (+) | cysteine and glycine-rich protein 2 | 113 |
| NM_175294 | 8430423A01R | Chr1:131762352-131784791 (+) | Nuclear ubiquitous casein and cyclin-dependent kinases substrate (JC7). | 99 |
| NM_144787 | Jmjd2c | Chr4:73303717-73467081 (+) | jumonji domain containing 2C | 185 |
| NM_016802 | Rhoa | Chr9:108375967-108407598 (+) | ras homolog gene family, member A | 97 |
| NM_148934 | Gtrgeo22 | Chr10:79791798-79798648 (+) | gene trap ROSA b-geo 22 | 94 |
| NM_009864 | Cdh1 | Chr8:105899006-105965884 (+) | cadherin 1 | 190 |
| NM_008602 | Miz1 | Chr18:77186924-77275519 (+) | Msx-interacting-zinc finger | 143 |
| NM_013827 | Mtf2 | Chr5:107136124-107178719 (+) | metal response element binding transcription factor 2 | 223 |
| NM_013512 | Epb4.114a | Chr18:34019351-34229942 (-) | erythrocyte protein band 4.1-like 4a | 179 |
| NM_145441 | Ubx4 | Chr12:4054764-4083225 (-) | UBX domain containing 4 | 133 |
| NM_020600 | Rps14 | Chr18:60999976-61003871 (+) | ribosomal protein S14 | 82 |
| NM_172860 | Cbfa2t2h | Chr2:153893456-153996294 (+) | core-binding factor, runt domain, alpha subunit 2, translocated to, 2 homolog (human) | 127 |
| NM_021878 | Jarid2 | Chr13:44305483-44495794 (+) | jumonji, AT rich interactive domain 2 | 146 |

Table 3. List of the first 50 hypertrapped genes.

For each gene it has been reported the identificative RefSeq, name, genomic localization, description, number of gene trapping experiments.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 14. Real-time RT-PCR verification of level of expression of hypertrapped genes.

The bar chart shows the levels of expressions of 10 hypertrapped genes (in red) and 10 genes trapped 1 or 2 times. 80% of hypertrapped genes are expressed at significantly higher levels than genes trapped at the median rate of 1 trap per gene.

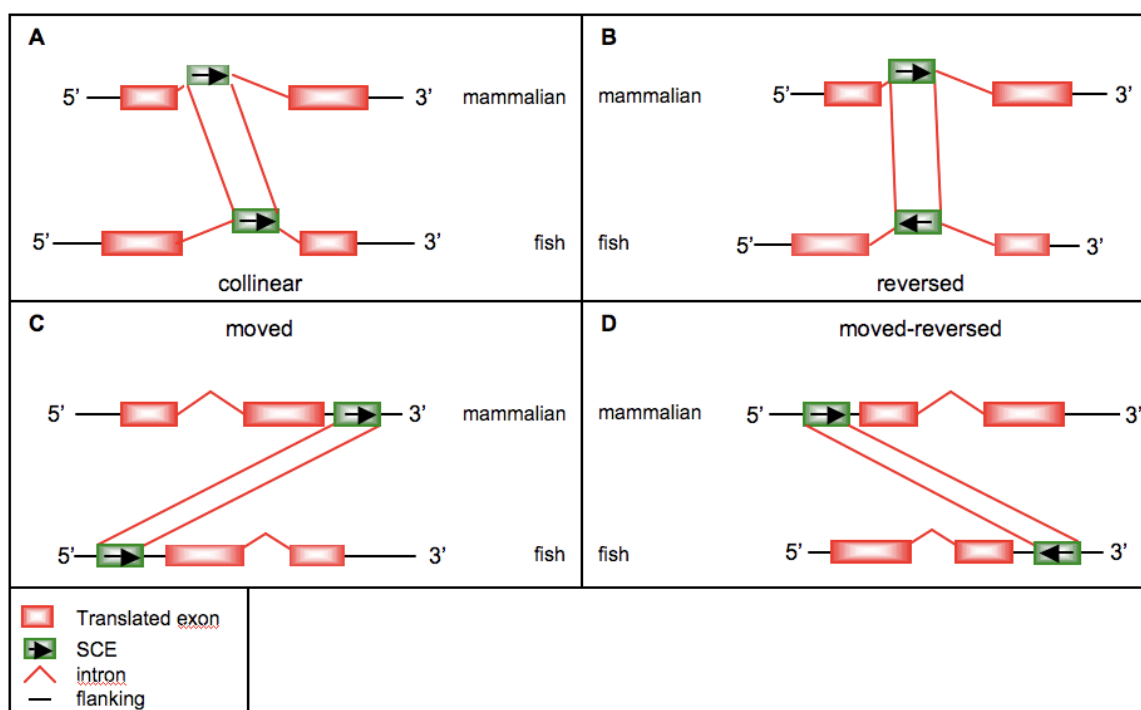


Figure 15. Shuffling categories of SCEs.

SCEs are categorized basing on their change in location and orientation in *Fugu rubiprens* with respect to their location and orientation in mouse locus. A) Collinear SCEs: elements that have not undergone any change in location or orientation within the entire gene locus. B) Reversed SCEs: elements that have changed their orientation in the fish locus with respect to the mouse locus, but have remained in the same portion of the locus. C) Moved SCEs: elements that have moved between the pre-gene, post-gene and intronic portion of the locus. D) Moved-reversed SCEs: elements that have undergone both of the above changes.

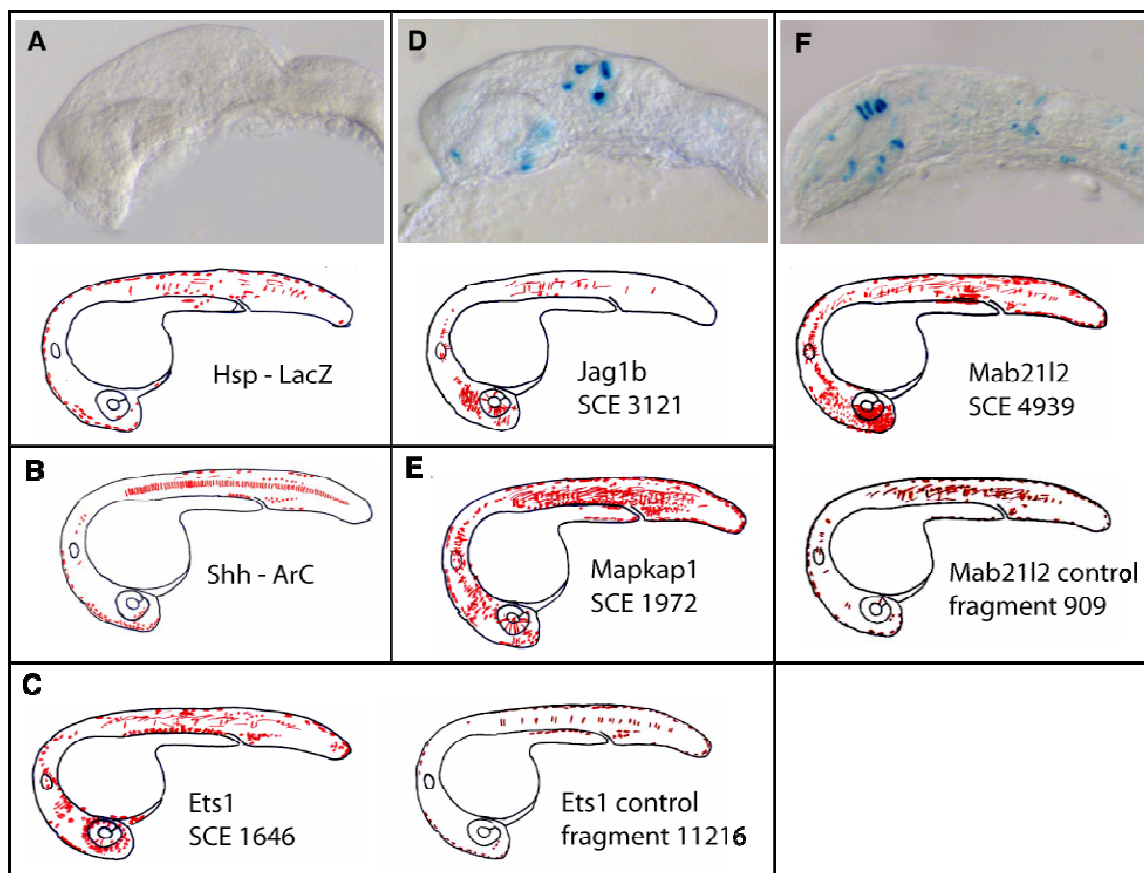


Figure 16. Expression profiles of X-Gal stained embryos.

A/B/C/D/E/F. Expression profiles of 1 day old X-Gal stained zebrafish embryos. Each expression map represents a composite overview of the LacZ positive cells of 65-175 embryos. Gene names and fragment/SCE are shown.

| gene | name | SCE length (bp) | Enhancer | n. of embryos | cells | muscle | notochord | CNS | eye | ear | vessels | other |
|-----------|-------|-----------------|----------|---------------|-------|-----------|-----------|-----------|----------|---------|----------|---------|
| no | lacZ | | neg | 161 | 40 | p-value | p-value | p-value | p-value | p-value | p-value | p-value |
| Shh | ArC | | pos | 96 | 242 | | 8.48E-07 | | | | | |
| Shh | 12058 | 45 | y | 139 | 69 | 6.86E-09 | | | | | | |
| Otx2 | 13988 | 51 | y | 111 | 93 | 0.6444 | | 0.006269 | 0.5536 | 0.3155 | | |
| Gata3 | 15402 | 40 | y | 107 | 103 | | | 0.398 | 0.5764 | 0.1906 | | 1 |
| Ets | 8744 | 40 | y | 105 | 180 | | | 0.002593 | | | 4.78E-06 | |
| Ets | 8745 | 46 | Y | 133 | 210 | | | 0.1558 | 0.6015 | 0.3619 | 2.15E-06 | |
| Ets | 8726 | 41 | Y | 159 | 345 | | | 0.05534 | 0.6131 | 0.1485 | 2.08E-06 | |
| Ets | 8728 | 48 | Y | 149 | 176 | | | 0.0444 | 0.129 | 0.07924 | 1.31E-05 | |
| Pax2b | 31027 | 39 | Y | 149 | 105 | | | 0.002374 | 0.06327 | 0.1902 | | |
| Pax6a | 15696 | 33 | Y | 133 | 122 | | | 8.21E-06 | 0.3343 | 0.01268 | | |
| Pax3 | 24781 | 42 | N | 124 | 67 | 0.02982 | | 0.5287 | | 1 | | |
| Zfpm2 | 23818 | 48 | Y | 140 | 119 | | | 1.49E-06 | 0.01296 | | 1 | |
| Zfpm2 | 23838 | 48 | Y | 131 | 148 | | | 0.0003376 | 0.04369 | 0.1231 | | |
| Tmeff2 | 26014 | 48 | N | 164 | 125 | | | 0.7654 | 0.2301 | 0.3371 | | 0.2801 |
| Tmeff2 | 26015 | 38 | Y | 120 | 159 | 0.001035 | | 0.303 | 0.2088 | | | |
| Tmeff2 | 26016 | 51 | Y | 109 | 148 | | | 0.0006306 | 0.0149 | 0.5862 | | |
| Jag1b | 16407 | 37 | N | 136 | 98 | 1 | | 0.1849 | | 1 | 1 | |
| Jag1b | 16408 | 55 | Y | 142 | 109 | | | 5.45E-08 | 0.006524 | 0.3245 | | |
| Jag1b | 16409 | 44 | N | 106 | 54 | 1 | | 0.5088 | 1 | 0.5058 | | |
| Mapkap1 | 17058 | 37 | Y | 143 | 295 | 0.6825 | | 0.05292 | 0.3788 | 0.6065 | | 1 |
| Mapkap1 | 17059 | 42 | Y | 136 | 171 | 0.6686 | | 0.004037 | 0.5973 | 0.077 | 0.5197 | |
| Mad211 | 23001 | 39 | Y | 142 | 317 | | | 1.24E-07 | 0.004985 | 0.2339 | | |
| Mad211 | 23002 | 37 | Y | 155 | 122 | | | 7.85E-08 | 0.004138 | | | |
| Hmx3 | 11669 | 150 | Y | 165 | 136 | | | 0.001029 | 0.07062 | 0.01423 | | |
| Lmx1b | 17027 | 300 | Y | 116 | 105 | | | 0.00762 | 0.1876 | 1 | | |
| 3110004 L | 5803 | 45 | N | 65 | 16 | 0.2929 | | | | | | 1 |
| 3110004 L | 5802 | 39 | Y | 122 | 320 | 0.1874 | 0.01209 | | | | | |
| Elmo1 | 6026 | 45 | Y | 103 | 76 | 0.007132 | 0.6848 | | | | | |
| Ets | 11216 | | N | 104 | 74 | 1 | | | | | | 0.6954 |
| Gata3 | 3255 | | N | 174 | 110 | 0.04481 | | 0.281 | 0.5739 | 0.02163 | | |
| 1300007 F | 2797 | | N | 157 | 115 | | | | | | | |
| Tmeff2 | 198 | | N | 145 | 23 | 0.7448 | | 0.6597 | | 0.3651 | | |
| Mad211 | 909 | | N | 165 | 92 | 0.06359 | | | 1 | 1 | 1 | |
| 3110004 L | 410 | | N | 107 | 23 | | | | | | | 0,0198 |
| Elmo1 | 10157 | | N | 146 | 38 | 0.287 | 0.8126 | | | | | |
| Shh | 11271 | | N | 165 | 83 | 3.34E-07 | | | 1 | 1 | 1 | |
| Impact | 5990 | | N | 150 | 101 | 0.6496 | | 0.2754 | | 0.0622 | | |
| Ubl7 | 268 | | N | 117 | 644 | 0.0003325 | | 7.15E-11 | 0,02555 | 0.6197 | | |
| Lmx1b | 11767 | | Y | 116 | 15 | 0.2743 | | | | 0.0707 | | 1 |
| Irx3 | 5945 | | N | 93 | 15 | 0.03938 | | | | | | |

Table 4. Analysis of X-Gal staining in zebrafish embryos co-injected with the Hsp promoter and SCEs or control fragments.

For each DNA fragment tested the following information is given: from left to right: the gene locus in which the DNA fragment is found, the size of the SCE, summary about the potentially enhancer function of the element (Y=yes, N=not), the number of embryos injected, the p-value indicating the significance of the number of cells observed in the fragment tested versus the LacZ: Hsp control for each tissue.

DISCUSSION

It might be surprising that about six years after the first draft of the human genome (Venter et al., 2001; Lander et al., 2001) and three years after the announcement of the completion of the genome sequence (International Human Genome Sequencing Consortium, 2004) we still do not have a complete set of all the genes that are encoded by the human genome. This is due to the ease with which sequence data is collected, and the difficulty in obtaining functional data in a similarly high throughput manner. It is for this reason that the functional characterization of every single gene within the mammalian genome is one of the major aims of the post-genomic era. Thus, in recent years, the interest in tools that enable genome-wide mutagenesis in a streamlined manner has increased significantly.

Among the available approaches used to identify and characterize novel genes, gene trapping in mouse ES cells has emerged as a powerful tool which enables analysis of mammalian gene function in a post-genomic era. The application of this technique in a genome-wide manner should allow the identification of most, if not all, active transcripts in the genome of ES cells and thus it was chosen as an innovative tool for genome annotation.

In our study, starting with a large data set of all available sequences derived from gene trapping experiments we investigated if they allowed us to decipher the ES cell transcriptome, as well as the mouse genome at a wider level. Notably, we found that 38% of trapclusters cannot be mapped to genomic regions previously annotated by other existing databases such as RefSeq, Ensembl, Fantom or Unigene (Fig. 17).

Moreover, we observed a richness in alternative splicing for 5' and 3' exons. Interestingly, using gene trapping, we refine the structure of existing known genes adding novel exons. The reason why we added a larger number of internal exons with respect to external ones may depend on the fact that this technique usually provides sequences from integration events which occur within introns. By RT-PCR experiments we were able to validate that 70% of these candidate exons are really expressed, and that often they exhibit a tissue specific pattern of expression. The identification of new exons on genes coming from a well-annotated database such as RefSeq stresses the incomplete annotation of these genes. These findings are in line with the fact that even though only a small portion of human genes is known to be lacking from computational predictions, the exact genomic structure of these genes seems to be correct only for approximately 50% of them.

We demonstrated with our findings that 40% of the exons that were added can be detected in ES cells, while a further 30% are expressed in a tissue specific manner and are not detectable in ES cells as was verified by testing four additional RNA sources. Thus it is reasonable hypothesize that a higher proportion of our novel exons could be verified if we investigated more developmental stages and tissues. Moreover, these results suggest that genes which are successfully trapped in ES cells are often expressed at very low levels in these cells, while in other tissues they could be expressed at higher levels thus, showing a specific pattern linked to specific tissues, stages and cell types upon differentiation.

The fact that gene trapping in ES cells could reveal a higher number of novel genes than it had been shown before using cDNA and EST based approaches is probably due to

the different levels of expression. Probably these genes are expressed at high levels at specific time points and in cell types (such as ES cells) that have not been used to produce libraries for EST collection.

RT-PCR, real-time PCR, in situ hybridization, as well as computational approaches (multispecies alignments, comparison with tiling array data) were employed to demonstrate that the 65% of our trapclusters correspond to novel genes which are truly expressed in ES cells. Particularly, the specific expression profile showed by in situ hybridization on the TCLG1417 predicted gene within the auditory pathway was interesting taking into consideration that it is a novel non-coding gene that does not fall in the much studied microRNA category.

It is well known that ES cells express a wide number of genes at basal level and a few hundreds genes at high levels (Sharov et al., 2003). One can therefore assume that highly expressed genes might be easier to identify using gene trapping techniques. Several studies indicated that the genome presents specific regions that are hot spots for this technique. In our work we demonstrated by real time PCR experiments that these hotspot regions correspond to genes which are significantly expressed in ES cells and that their expression levels are comparable to Oct4 gene, a well known ES cell marker. Hypertrapped and trapped categories both contain genes that are related to basic molecular functions of the cell, such as transcription, translation and degradation of proteins. Hypertrapped genes show a balanced subselection of the same types of genes. Interestingly the latter dataset includes some factors that are involved in the early stages of development such as *Erdr1*, *Klf9*, *Gabarapl2*, *Rbpms*, *Cfdp1*. These factors might be highly expressed in pluripotent cells to be “ready to go” once these cells differentiate into

a specific fate (i.e. cell type). It could be hypothesized that their expression might be under the control of transcription factors which are known to guarantee the maintenance of pluripotency in the germinal cell, such as for example Oct4, Nanog, Sox2 and STAT3.

A comparative analysis of hypertrapped genes against already known sets of genes involved in the “stemness profile”, derived from expression profiling experiments (Vogel et al., 2003) showed a remarkably low overlap. In particular it is striking that genes belonging to our set of hypertrapped genes, which are expressed at significantly higher levels than “normally trapped” genes (as demonstrated by real-time PCR) are not present in the datasets published. This result, together with that obtained on Oct4, suggests that hypertrapped genes might identify a set of genes whose expression is tightly controlled, as it is for Oct4, and which could thus be difficult to identify by expression profiling, while playing an important role in the biology of these cells.

Our data taken together indicates clearly that gene trapping in ES cells holds high value for biology and that its utility extends far beyond its use as a mere mutagenesis tool. We demonstrated that thousands of novel genes and transcripts exist which had never been annotated; thus we can conclude that gene-trap mutagenesis is an efficient approach for annotating and dissecting the function of mammalian genes.

Another fascinating challenge of the post genomic era is that of understanding the intricate processes of gene regulation in vertebrates. Comparative genomics is one of the approaches commonly used to identify non-coding regions of the genome which are conserved across evolution and which might play a role in the process of gene regulation. In order to define novel putative regulatory elements in the vertebrate genome we focused our attention on the conservation of non-coding elements between fish and

mammalian genomes. We hypothesized that over such long evolutionary distances, over which entire genes are known to shuffle, non-coding elements should shuffle at even greater extent given that their function is often position and orientation independent. Thus we developed a pipeline using the CHAOS algorithm for detecting conserved elements which could have shuffled across evolution and we then proceeded to test a subset of them (as well as a set of matched negative controls) using an enhancer assay in zebrafish to investigate their functionality.

Using a restricted set of candidate non-coding regulatory sequences identified by comparative analysis we were able to demonstrate their *cis*-acting regulatory activity in transgenic zebrafish. We demonstrated by co-injection the enhancer activity of the majority (80%) of the elements identified. We followed the expression profile of each fragment in 24hrs old co-injected embryos. As a positive control ArC expression was verified in the notochord as previously reported. The fact that these elements are well conserved in *Fugu*, demonstrates that the expression regulation of expression of genes involved in development is conserved. The transient expression of these elements in zebrafish showed an interesting tissue specific pattern for most of them, and where the pattern of the neighbouring gene was known the pattern produced by the patten was often similar. Notably, our data demonstrated that 80% of the elements tested do enhance transcription in vivo as compared to a single element in the control set of fragments, and that most drive tissue specific expression of a minimal promoter.

Taken together our data demonstrates that the combination of a comparative genomics approach and functional screening is able to produce a large data set that will be useful for further investigation helping to expand the understanding of the genome and

understanding the intricate mechanisms of gene regulation, taking into account novel and as yet not very well understood players such as those hiding in the non-coding realms of the genome.

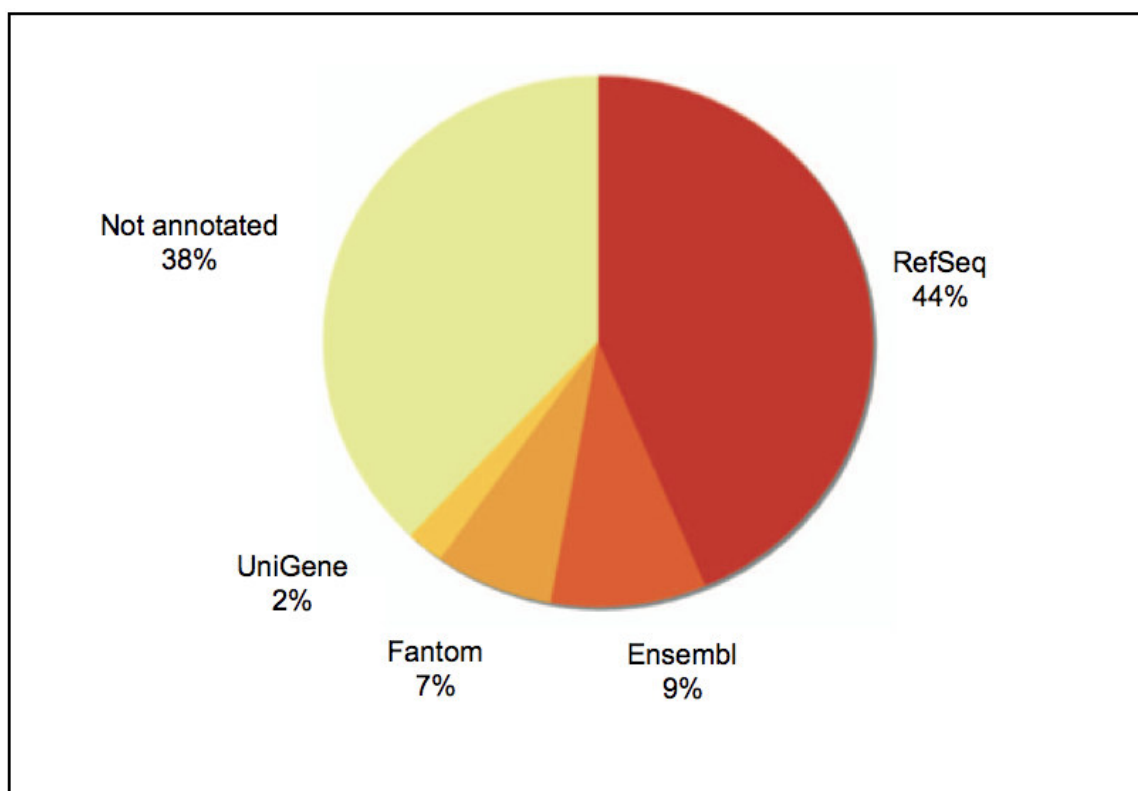


Figure 17. Annotation of Trap Clusters.

Pie chart showing the annotation of genomic regions mapped by trapclusters, indicating that 38% of the trapclusters analyzed cannot be mapped to regions of the genome which have already been annotated with gene structures by RefSeq, Ensembl, Fantom or UniGene.

REFERENCES

Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J, Gorrell JH, Chinault AC, Belmont JW, Miller W, Gibbs RA. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* 8 (1): 29-40.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. (2002) Whole genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science.* 297 (5585): 1301-1310.

Bagheri-Fam S, Ferraz C, Demaille J, Sherer G, Pfeifer D. (2001) Comparative genomics of the SOX9 region in human and *FUGU rubripes*: conservation of short regulatory sequence elements within large intergenic regions. *Genomics.* 78: 73-82.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. (2004) Ultraconserved elements in the human genome. *Science.* 304 (5675): 1321-1325.

Bell AC, West AG, and Felsenfeld G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic. *Science.* 291: 447-450.

Bellen HJ, et al. (1996) P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* 3: 1288-1300.

Birney E. Ensembl 2005 *Nucleic Acids Res.* (2005) Jan 1; 33 Database issue: D447-D453.

Blakwood EM, Kadonaga JT. (1998) Going the distance: a current view of enhancer action. *Science.* 281 (5373): 60-63.

Boffelli D, Nobrega MA, Rubin EM. (2005) Comparative genomics at the vertebrate extremes. *Nat Rev Genet.* 5: 456-465.

Boheler KR, Stern MD. (2003) The new role of SAGE in gene discovery. *Trends Biotechnol.* 21: 55-57.

Bonaldo P, Chowdhury K, Stoykova A, Torres M, Gruss P. (1998) Efficient gene trap screening for novel developmental genes using IRES β geo vector and in vitro preselection. *Exp Cell Res.* 244: 125-136.

Brandl S. (2002) Antisense-RNA regulation and RNA interference. *Biochim Biophys Acta.* 1575: 15-25.

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S. (1993) Characterization of the pufferfish (FUGU) genome as a compact model vertebrate genome. *Nature*. 366(6452): 265-268.

Burns CE and Zon LI. (2002) Portrait of a stem cell. *Dev Cell*. 3: 612-613.

Carnici P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N Oyama R, Ravassi T, Lenthard B, Wells C, et al. FANTOM consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005) The transcriptional landscape of the mammalian genome. *Science*. 309: 1559-1563.

Carnici P, Waki K, Shiraki T, Konno H, Shibata K et al. (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res*. 13: 1273-1289.

Carter D, Chakalowa L, Osborne CS, Dai YF, Fraser P. (2002) Long-range chromatin regulatory interaction *in vivo*. *Nat Genet*. 32: 623-626.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*. 308: 1149-1154.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigò R, Gingeras TR, Antonarakis SE, Reymond A. (2007) Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res*. 17 (6): 746-759.

Dermitzakis ET, Kirkness E, Schwartz S, Birney E, Reymond A, Antonarakis SE. (2004) Comparison of human chromosome 21 conserved non genic sequences (CNGs) with mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res*. 5: 852-859

Dermitzakis ET, Reymond A, Antonarakis SE. (2005) Conserved non-genic sequences -an unexpected feature of mammalian genomes. *Nat Rev Genet*. 6(2): 151-157.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*. 302: 1033-1035.

Dermitzakis ET, Stranger BE. (2006) Genetic variation in human gene expression. *Hum genomics*. 2 (6): 383-390.

- De-Zolt S, Schnutgen F, Seisenberger C, Hansen J, Hollatz M, Floss T, Ruiz P, Wurst W, von Melchner H. (2006) High-throughput trapping of secretory pathway genes in mouse embryonic stem cells. *Nucl Acid Res.* 34.
- Dikmeis T, Plessy C, Rastegar S, Aanstad P, Herwig R et al. (2004) Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in zebrafish embryo. *Genome Res.* 14: 228-238.
- Dorsett D. (1999) Distant liaisons: long range enhancer-promoter interactions in *Drosophila*. *Curr Opin Genet Dev.* 9(5): 505-514.
- Down TA, and Hubbard TJ. (2002) Computational deletion and location of transcription start sites in mammalian genomic DNA. *Genome Res* 12: 458-462.
- Dufresne M, Hua-Van A, El Wahab HA, Ben M'Barek S, Vasnier C, Teyssst L, Kema GH, Daboussi MJ. (2007) Transposition of a fungal miniature inverted-repeated transposable element through the action of a Tc1-like transposase. *Genetics.* 175(1): 441-452.
- Durick K, Mendlein J, Xanthopoulos KG. (1999) Hunting with traps: genome-wide strategies for gene discovery and functional analysis. *Genome Res.* 9: 1019-1025.
- Elnitski L, Miller W, Hardison R. (1997) Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region. Role of basic helix-loop-helix proteins. *J. Biol Chem.* 272(1): 369-378.
- Ewing B, Green P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet.* 25: 232-234.
- Fantom Consortium and the Riken exploration Research Group phase I and II team. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* 420: 563-573.
- Feshotte C, Zhang X, Wessler S. (2002) Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons, in N.L. Craig, Craigie R, Gellert M, Lambowitz A M, (eds.), *Mobile DNA II*, pp. 1147-1158 ASM Press, Washington DC.
- Finegan JA, Bartleman B, Wong PY. (1989). A window for the study of prenatal sex hormone influences on postnatal development. *J Genet Psychol.* 150(1): 101-112.
- Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR. (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* 13(3): 341-346.

- Friedel RH, Plump A, Lu X, Spilker K, Jolicoeur C, Wong K, Venkatesh T R, Yaron A, Hynes M, Chen B, Okada A, McConnel SK, Rayburn H, Tessier-Lavigne M. (2005) gene targeting using a promoterless gene trap vector ("targeted trapping") is an efficient method to mutate a large fraction of genes. *PNAS*. 102(37): 13188-13193.
- Friedrich G, Soriano P. (1991). Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice *Genes Dev*. 5:1513-1523.
- Gerber WV, Yatskievych TA, Antin PB, Correia KM, Conlon RA, Krieg PA. (1999) The RNA binding protein, hermes, is expressed at high levels in the developing heart. *Mech Dev*. 80: 77-86.
- Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, et al. (2003) Regulatory roles of conserved intergenic domains in vertebrate *Dlx* *Bigene* clusters. *Genome Res*. 13: 533-543.
- Gondo Y, Okada T, Matsuyama N, Saitoh Y, Yanagisawa Y, Ikeda JE. (1998) Human megasatellite DNA RS 447: copy number polymorphism and interspecies conservation. *Genomics*. 54(1): 39-49.
- Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T Bock JH, Slightom JL, Goodman M. (1996) Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol Phylogenet Evol*. 5(1): 18-32.
- Hansen J, et al. (2003) A large scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proc Natl Acad Sci USA*. 100: 9918-9922.
- Hardison RC. (2000) Conserved non-coding sequences are reliable guides to regulatory elements. *Trends genet*. 16(9): 369-372.
- Hayashizaki Y and Carnici P. (2006). Genome network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet*. 2: 492- 497.
- Hirotsune S. (2003) An expressed pseudogene regulates mRNA stability of its homologous coding gene. *Tanpakushitsu KaKusan Koso*. 4(14): 1908-1912.
- Hutvagner G, Zamore PD. (2002) RNAi nature abhors a double strand. *Curr Opin Genet Dev*. 12: 225-232.
- International Human genome Sequencing Consortium: finishing the euchromatic sequence of the human genome. (2004) *Nature*. 431: 931-945.
- International Human genome Sequencing Consortium: initial sequencing and analysis of the human genome. (2001) *Nature*. 409: 520-562.

- Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, et al. (2002) A stem cell molecular signature. *Science*. 298: 601- 604.
- Jurka J. (1998) Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol*. 8(3): 333-337.
- Kammandel B, Chowdhury K, Stoykova A, Aparicio S, Brenner S, et al. (1999) Distinct cis –essential modules direct the time-space pattern of the Pax6 gene activity. *Dev Biol*. 205: 79- 97.
- Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M et al. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*. 409: 685-690.
- Kimura-Yoshida C, Kitajima K, Oda-Ishii I, Tian E, Suzuki M, Yamamoto M, Suzuki T, Kobayashi M, Aizawa S, Matsuo I. (2004) Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*. 131(1): 57-71.
- Kudoh T, Dawid IB. (2001) Zebrafish mab21l2 is specifically expressed in the presumptive eye and tectum from early somitogenesis onwards. *Mech Dev*. 109(1): 95-98.
- Kudoh T, Tsang M, Hukriede NA, Chen X, Dedekian M, Clarke CJ, Kiang A, Schultz S, Epstein JA, Toyama R et al. (2001) A gene expression screen in zebrafish embryogenesis. *Genome Res*. 11(12): 1979-1987.
- Kumar S, Hedges SB. (1998) A molecular timescale for vertebrate evolution. *Nature*. 392: 917- 920.
- Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. *Nature*. 409: 860- 921.
- Lee TH, Streb JW, Georer MA, Miano JM. (2006) Tissue expression of the novel serine carboxypeptidase Scep1. *J Histochem Cytochem*. 54: 701-711.
- Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, et al. (2003) A long range Shh enhancer regulates expression in the developing limb and fin and disassociated with preaxial polydactyly. *Hum Mol Genet*. 12: 1725-1735.
- Liang F, Holt I, Pertea G, Karamysheva S, Salzberg SL, Quachenbush J. (2000) Gene index analysis on the human genome estimates approximately 120,000 genes. *Nat Genet*. 25: 239-240.
- Liang K, Lin Y, Zhang Y, Chen Z, Zhang P, Zang H. (2004) Developmental expression of amphioxus GABAA receptor-associated protein-like 2 gene. *Dev Genes Evol*. 214: 339-341.

- Liu J, Gough J, Rost B. (2006) Distinguishing protein-coding from non-coding mRNAs through support vector machines. *PLoS Genet.* 2.
- Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, and Frazer KA. (2000) Identification of a coordinate regulator for interleukins 4, 13, and 5 by cross-species comparisons. *Science.* 288: 136-140
- Lynch S, Kewalramani A. (2003) Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol Biol Evol.* 20: 563-571.
- Marshall H, Studer M, Propperl H, Aparicio S, Kuroiwa A et al. (1994) A conserved retinoic acid response element required for early expression of homeobox gene Hoxb-1. *Nature* 370: 567-571.
- Martin KM Metcalfe JC Kemp PR. (2001) Expression of Klf9 and Klf13 in mouse development *Mech Dev.* 103(1-2): 149-151.
- McEwen GK, Woolfe A Goode D, Vavouri T, Callaway H et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates : a genomic and functional analysis. *Genome Res.* 16: 451-465.
- Miller W, Makova KD, Nekrutenko A, Hardison RC. (2004) Comparative genomics. *Annu Rev Genomics Hum Genet.* 5: 15-56.
- Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature.* 420: 520-562.
- Mukhopadhyay P, Greene RM, Pisano MM. (2006) Expression profiling of transforming growth factor beta superfamily genes in developing orofacial tissue. *Birth Defects Res A Clin Mol Teratol.* 76: 528-543.
- Muller F, Williams DW, Kobolak J, Gauvry L, Goldspink G, Orban L, Maclean N. (1997) Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol Reprod Dev.* 47(4): 404-412.
- Muller F, Chang B, Albert S, Fisher N, Tora, L, Strahle U. (1999) Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. *Development.* 126(10): 2103-2116.
- Nagano K, Taoka M, Yamauchi Y, Itagaki C, Shinkawa T, Nunomura K, Okamura N, Takahashi N, Izumi T, Isobe T. (2005) Large-scale identification of proteins expressed in mouse embryonic stem cells. *Proteomics.* 5(5): 1346-1361.
- Neumann B, Kubika P Barlow DP. (1995) Characteristics of imprinted genes. *Nat Genet.* 9(1): 12-13.

- Niwa H, Miyazaki J, Smith AG. (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet.* 24: 372-376.
- Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. (2004) Megabase deletions of gene deserts result in viable mice. *Nature.* 431 (7011): 988-993.
- Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. (2003) Scanning human gene deserts for long-range enhancers. *Science.* 302: 413.
- Nobrega MA, Pennacchio LA. (2004) Comparative genomic analysis as a tool for biological discovery *J Physiol.* 554: 31-39.
- Oberwinkler J, Lis A, Giehl KM, Flockerzi V, Philipp SE. (2005) Alternative splicing switches the divalent cation selectivity of TRPM3 channels. *J Biol Chem.*
- O'Brien SJ et al. (1999) The promise of comparative genomics in mammals. *Science.* 286: 458-462, 479-481.
- Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* 420: 563-573.
- Orgel LE, Crick FH, Sapienza C. (1980) Selfish DNA. *Nature.* 288 5792: 645-646.
- Ota T, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K., Kimura K, et al. (2004) Complete sequencing and characterization of 21,234 full-length human cDNAs. *Nat Genet.* 36: 40-45.
- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA et al. (2006) *In vivo* enhancer analysis on human conserved non-coding sequences. *Nature.* 444: 499-502.
- Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. (2002) "Stemness": Transcriptional profiling of embryonic and adult stem cells. *Science.* 298: 597-600.
- Rowich DH, Echelard Y, Danielian PS, Gellner K, Brenner S, et al. (1998) Identification of an evolutionary conserved 110 base-pair cis-acting regulatory sequence that govern Wnt-1 expression in the murine neural plate. *Development.* 125:1735-1746.
- Russel LB, Montgomery CS, and Raymer GD. (1982) Analysis of the albino-locus region of the mouse, IV: characterization of 34 deficiencies. *Genetics.* 100: 427-453.

- Santagati F, Abe K, Shmidt V, Shmitt-John T, Suzuki M et al. (2003) Identification of cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved syntenies. *Genetics*. 165: 235-242.
- Schuler GD, et al. (1996) A gene map of the human genome. *Science*. 274: 540-546.
- Schuler GD. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med*. 75: 694-8.
- Semplici F, Meggio F, Pinna LA, Oliviero S. (2002) CK2-dependent phosphorylation of the E2 ubiquitin conjugating enzyme UBC3B induces its integration with beta-TrCP and enhances beta-catenin degradation. *Oncogene*. 21: 3978-3987.
- Sharov AA, Piao Y, Matoba R, Dudekula DB, Qian Y, et al. (2003) Transcriptome Analysis of Mouse Stem Cells and Early Embryos. *PLoS Biol*. 1(3): e74.
- Siddal NA, McLaughlin EA, Marriner NL, Hime GR. (2006) The RNA binding protein Musashi is required intrinsically to maintain stem cell identity. *Proc natl Acad Sci USA*. 103: 8402-8407.
- Skarnes WC, Auerbach BA, Joyner AL. (1992) A gene trap approach in mouse embryonic stem cells the LacZ reporter is activated by splicing, reflects the endogenous gene expression, and is mutagenic in mice. *Genes Dev*. 6: 903-918.
- Skarnes WC, Moss JE, Hurtley SM, Beddington RS. (1995) Capturing genes encoding membrane and secreted proteins important for mouse development. *Proc Natl acad Sci USA*. 92: 6592-6596.
- Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, Young SG, Ruiz P, Soriano P, Tessier-Lavigne M, Conklin BR, Stanford WL, Rossant J, International Gene Trap Consortium (2004) A public gene trap resource for mouse functional genomics. *Nat Genet*. 36(6): 543-544.
- Spitz F, Gonzales F, Duboule D. (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*. 113: 405-417.
- Sprague J, Clements D, Conlin T, Edwards P., Frazer K, Shaper K, Segerdell E, Song P, Sprunger B, Westerfield M. (2003) The Zebrafish International Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res*. 31(1): 241-243.
- Stanford WL, Cohn JB, and Cordes SP. (2001) Gene-trap mutagenesis : past, present and beyond. *Nature Rev Genet*. 2: 756-768.
- Tanaka TS, Kunath T, Kimber WL, Jaradat SA, Stagg CA, et al. (2002) Gene expression profiling of embryo-derived stem cells reveals candidate genes associated with pluripotency and lineage specificity. *Genome Res*. 12: 1921-1928.

The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 306: 636-640.

The ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. (in press).

The zebrafish International Resource Center (<http://zfin.org/>).

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analysis of multi-species sequences from targeted genomic regions. *Nature*. 424: 788-793.

Trinklein ND, Karaoz U, Wu J, Halees A, Force Aldred S, Collins PJ, Zheng D, Zhang ZD, Gerstein MB, Snyder M, Myers RM, Weng Z. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res*. 17(6): 720-731.

Vavuri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. (2006) Defining a genomic radium for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet*. 22: 5-10.

Venter et al. (2001) The sequence of the human genome *Science*. 291(5507): 1304-51.

Vogel G. (2003) Stem cells. 'Stemness' genes still elusive. *Science*. 302: 371.

Von Melchner H, De Gregori JV, Rayburn H, Reddy S, Friedel C, Ruley HE. (1992) Selective disruption of genes expressed in totipotent embryonal stem cells. *Genes Dev*. 6(6): 919-927.

Walton RZ, Bruce AE, Olivey HE, Najib K, Johnsons V, Earley JU, Ho RK, Svensson EC. (2006) *Fog1* is required for cardiac looping in zebrafish. *Dev Biol*. 289(2): 482-493.

Waterson RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*. 420: 520- 562.

Westerfield M, Wegener J, Jegalian BG, De Robertis EM, and Püshel AW. (1992) Specific activation of mammalian Hox promoters in mosaic transgenic zebrafish. *Genes Dev*. 6: 591-598.

Woolfe A, Goodson DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *Plos Biol*. 3(1): e7.

- Xi Q, He W, Zhang XH, Le HV Massagué J. (2007) Genome-wide impact of the BRG1 SWI/SNF chromatin remodeler on the TGFbeta transcriptional program. *J Biol Chem.*
- Yamaguchi K, Hidema S, Mizuno S. (1998) Chicken chromobox proteins: cDNA cloning of CHCB1, -2, -3 and their relation to W-heterochromatin *Exp Cell Res.* 242: 303-314.
- Zako T, Iizuka R, Okochi M. Nomura T, Ueno T, Tadakuma H, Yohda M, Funatsu T. (2005) Facilitated release of substrate protein from prefoldin chaperonin. *FEBS Lett.* 579: 3718-3724.
- Zambrowicz BP, Friedrich GA, Buxton EC, Lilleberg SL, Person C, and Sands AT. (1998) Disruption and sequence identification of 2.000 genes in mouse embryonic stem cells. *Nature.* 392: 608-611.
- Zecchin E, Conigliaro A, Tiso N, Argenton F, Bortolussi M. (2005) Expression analysis of jagged genes in zebrafish embryos. *Dev Dyn.* 233: 638-645.
- Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein M. (2007). Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res.*
- Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, et al. (2005) Integrated pseudogene annotation for chromosome 22: Evidence for transcription. *J Mol Biol.* 349: 27-45.

Publications arising from this thesis

Roma G.¹, Cobellis G.¹, **Claudiani P.**¹, Maione F., Cruz P., Tripoli G., Sadiello M. Peluso I. and Stupka E. A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. *Genome Research*, 2007.

¹ These authors equally contributed at the work.

Sanges R., Kalmar E., **Claudiani P.**, D'Amato M. Muller F. and Stupka E. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biology*, 2006.

APPENDIX: PRIMERS USED

Primers used to amplify novel transcripts

| forward_name | forward_seq |
|--------------|---------------------------|
| FP81.1.2 | GAGCCCTGCTGTGGGTGAAGAACT |
| FP81.1.2 | GAGCCCTGCTGTGGGTGAAGAACT |
| FP81.2.2 | TTGTTCCAGAAGGAGGCACAGTCC |
| FP330.1.1 | CACGTCACATCACTGCTCCCAACT |
| FP330.1.1 | CACGTCACATCACTGCTCCCAACT |
| FP330.2.2 | GAGGGGTTCTTGCCTGTTTGTGTG |
| FP330.2.2 | GAGGGGTTCTTGCCTGTTTGTGTG |
| FP400.1.2 | TGGGAAGTTGAGCAGGAAACTCCA |
| FP400.1.2 | TGGGAAGTTGAGCAGGAAACTCCA |
| FP400.2.2 | GTCCTGGTCCCAAGACCTCAGCTT |
| FP400.2.2 | GTCCTGGTCCCAAGACCTCAGCTT |
| FP455.1.2 | GCCTACTGCCTCGTTCCCAGTTTC |
| FP455.1.2 | GCCTACTGCCTCGTTCCCAGTTTC |
| FP455.2.4 | AGCCCAGAGAGACAGACCGACAAG |
| FP467.1.1 | TTGCTGCGGAGTGTCTCTGAATTG |
| FP467.1.1 | TTGCTGCGGAGTGTCTCTGAATTG |
| FP467.1.1 | TTGCTGCGGAGTGTCTCTGAATTG |
| FP467.2.1 | ATTTGAAGCTGCCCCCTCAAAGGAA |
| FP467.2.1 | ATTTGAAGCTGCCCCCTCAAAGGAA |
| FP467.2.1 | ATTTGAAGCTGCCCCCTCAAAGGAA |
| FP467.3.1 | CCCTAGTTGCCAGAAATTGCAGA |
| FP467.3.1 | CCCTAGTTGCCAGAAATTGCAGA |
| FP486.1.2 | TCAGGGCATGGAGCAAATCTTCTG |
| FP486.1.2 | TCAGGGCATGGAGCAAATCTTCTG |
| FP486.1.2 | TCAGGGCATGGAGCAAATCTTCTG |
| FP486.2.2 | TTCGTGCTTGAGATGCAGAGGGTA |
| FP486.2.2 | TTCGTGCTTGAGATGCAGAGGGTA |
| FP486.2.2 | TTCGTGCTTGAGATGCAGAGGGTA |
| FP486.3.2 | CCCATGTCTTGTGGGGACAAAGAG |
| FP486.3.2 | CCCATGTCTTGTGGGGACAAAGAG |
| FP724.1.1 | CCTCTCGGAAAAAGGGTCAACTGG |
| FP724.1.1 | CCTCTCGGAAAAAGGGTCAACTGG |
| FP724.2.2 | CAGCCTGCTAGGATGCCTCTGTTG |
| FP724.2.2 | CAGCCTGCTAGGATGCCTCTGTTG |
| FP757.1.3 | ATCCCTGAGGAGCTGACGGTGAAC |
| FP757.1.3 | ATCCCTGAGGAGCTGACGGTGAAC |

| | |
|------------|--------------------------|
| FP757.2.3 | GTGCTTTGTTTCGCAGGCATTTTC |
| FP757.2.3 | GTGCTTTGTTTCGCAGGCATTTTC |
| FP869.1.1 | AGCCAGCTTCTCTCACCCTTGGA |
| FP869.1.1 | AGCCAGCTTCTCTCACCCTTGGA |
| FP869.2.2 | ACCGTGGATGAGGAGATCGATGAA |
| FP947.1.2 | CCCTAAGCGAACCTTGGAGAATGC |
| FP947.1.2 | CCCTAAGCGAACCTTGGAGAATGC |
| FP947.2.3 | CCATTGAGCCACCATCCACATACA |
| FP947.2.3 | CCATTGAGCCACCATCCACATACA |
| FP978.1.3 | AAGGAGAAAGCCCCTTCCTCGAA |
| FP978.1.3 | AAGGAGAAAGCCCCTTCCTCGAA |
| FP978.2.2 | ACAGCCTGGGAAAATGGAGATGCT |
| FP1004.1.3 | GGAGCCGGTGACACTGAATAGCAC |
| FP1004.1.3 | GGAGCCGGTGACACTGAATAGCAC |
| FP1004.1.3 | GGAGCCGGTGACACTGAATAGCAC |
| FP1004.2.2 | GCACAAGGGTGGCTGATTCAAGAC |
| FP1004.2.2 | GCACAAGGGTGGCTGATTCAAGAC |
| FP1004.2.2 | GCACAAGGGTGGCTGATTCAAGAC |
| FP1004.3.1 | CCTCACCATATCGGCCCTTTCCTA |
| FP1004.3.1 | CCTCACCATATCGGCCCTTTCCTA |
| FP1004.3.1 | CCTCACCATATCGGCCCTTTCCTA |
| FP1113.1.2 | CAGTTGTCTGATGGGGGACTGAGA |
| FP1113.1.2 | CAGTTGTCTGATGGGGGACTGAGA |
| FP1113.1.2 | CAGTTGTCTGATGGGGGACTGAGA |
| FP1113.2.1 | TGCTGTTAATAATGGGCCCTCCA |
| FP1113.2.1 | TGCTGTTAATAATGGGCCCTCCA |
| FP1113.2.1 | TGCTGTTAATAATGGGCCCTCCA |
| FP1113.3.1 | ATATGGCTGCTCCACTTCCCCAGT |
| FP1113.3.1 | ATATGGCTGCTCCACTTCCCCAGT |
| FP1131.1.1 | TGAGGCAATTCAGGGGAGAAAACA |
| FP1131.1.1 | TGAGGCAATTCAGGGGAGAAAACA |
| FP1131.2.2 | CACCCCTCCCAGCCTTAGAGAAGA |
| FP1153.1.1 | ACGGAAACTGGCATCTGCAAGAAA |
| FP1153.1.1 | ACGGAAACTGGCATCTGCAAGAAA |
| FP1153.2.3 | GAACAAGCCAAAACCCTGGGAGAG |
| FP1153.2.3 | GAACAAGCCAAAACCCTGGGAGAG |
| FP1205.1.1 | CTTGGGGTGGAGCACGAATGTAAG |
| FP1259.1.1 | TCCTTGCTACCCCGGATTTCATTC |
| FP1259.1.1 | TCCTTGCTACCCCGGATTTCATTC |
| FP1259.2.1 | TGACGTGGGAGAGAATGTGAGTGC |
| FP1450.1.4 | CGCGATGCTGTTCTGTGATTCT |
| FP1450.1.4 | CGCGATGCTGTTCTGTGATTCT |
| FP1450.2.1 | GTTTCTCACGAGATGCTGCCCTTC |

| | |
|------------|---------------------------|
| FP1520.1.3 | TTCCTGCTCCACATGGTGTTTCTG |
| FP1520.1.3 | TTCCTGCTCCACATGGTGTTTCTG |
| FP1520.2.3 | TTGACTCAGGTGAGGGCCTAGGTG |
| FP1520.2.3 | TTGACTCAGGTGAGGGCCTAGGTG |
| FP1541.1.2 | GTCATCAGCTTCGTGACTGGGTGA |
| FP1541.1.2 | GTCATCAGCTTCGTGACTGGGTGA |
| FP1541.2.4 | GGACCACCAGTGGATTCCCTCTGT |
| FP1581.1.5 | CGGCTTTGGAAATACGAACTTGGA |
| FP1581.1.5 | CGGCTTTGGAAATACGAACTTGGA |
| FP1581.1.5 | CGGCTTTGGAAATACGAACTTGGA |
| FP1581.2.2 | TGGGTCACATATCTGGGGAGGTGT |
| FP1581.2.2 | TGGGTCACATATCTGGGGAGGTGT |
| FP1581.2.2 | TGGGTCACATATCTGGGGAGGTGT |
| FP1581.3.1 | CCCTTCTGATAGCATCTTCCTCTGA |
| FP1581.3.1 | CCCTTCTGATAGCATCTTCCTCTGA |
| FP1590.1.1 | TCAGCTAATGGCATAGGGCTTCCA |
| FP1590.1.1 | TCAGCTAATGGCATAGGGCTTCCA |
| FP1590.2.4 | AAAAGCCCGATCACCACAGCTTCT |
| FP1647.1.1 | AAGGACTCGAACCAGCGAATCCAG |
| FP1647.1.1 | AAGGACTCGAACCAGCGAATCCAG |
| FP1647.1.1 | AAGGACTCGAACCAGCGAATCCAG |
| FP1647.1.1 | AAGGACTCGAACCAGCGAATCCAG |
| FP1647.2.1 | CACGTCAGTTTGGCTTCATTGTGC |
| FP1647.2.1 | CACGTCAGTTTGGCTTCATTGTGC |
| FP1647.2.1 | CACGTCAGTTTGGCTTCATTGTGC |
| FP1647.2.1 | CACGTCAGTTTGGCTTCATTGTGC |
| FP1647.3.5 | TTTGGACACCACACAAGGTGATGC |
| FP1647.3.5 | TTTGGACACCACACAAGGTGATGC |
| FP1647.3.5 | TTTGGACACCACACAAGGTGATGC |
| FP1647.3.5 | TTTGGACACCACACAAGGTGATGC |
| FP1647.4.4 | GGGTCATCTCTTCCAATCCAGTGC |
| FP1647.4.4 | GGGTCATCTCTTCCAATCCAGTGC |
| FP1647.4.4 | GGGTCATCTCTTCCAATCCAGTGC |
| FP1688.1.1 | CACTCAGCTTTCTACGGCCCCTCT |
| FP1688.1.1 | CACTCAGCTTTCTACGGCCCCTCT |
| FP1688.2.1 | TGACCAACGGAAGGAGGAACACAT |
| FP1753.1.1 | TCCGCAGCACTTCCCATCTGTTAT |
| FP1753.1.1 | TCCGCAGCACTTCCCATCTGTTAT |
| FP1753.1.1 | TCCGCAGCACTTCCCATCTGTTAT |
| FP1753.2.3 | TGCCTGTGCAGTCCTTACTCAACG |
| FP1753.2.3 | TGCCTGTGCAGTCCTTACTCAACG |
| FP1753.2.3 | TGCCTGTGCAGTCCTTACTCAACG |
| FP1753.3.2 | AGTGTGCCTTGTGCTGTTGTCCAG |

| | |
|------------|---------------------------|
| FP1753.3.2 | AGTGTGCCTTGTGCTGTTGTCCAG |
| FP1777.1.2 | GCACTGAAAGCCCCTGATTGAAGA |
| FP1777.1.2 | GCACTGAAAGCCCCTGATTGAAGA |
| FP1777.2.2 | AAGCCGAGTATTGTGGGTGTGGAA |
| FP1928.1.1 | G TTCAGGTGAGCCGAGAGCAGTGT |
| FP1928.1.1 | G TTCAGGTGAGCCGAGAGCAGTGT |
| FP1928.2.1 | TGGGATGTTGCTTGATGACACCAC |
| FP1928.2.1 | TGGGATGTTGCTTGATGACACCAC |
| FP2005.1.2 | CGTGGCTCCCTCTACCAATTCTCC |
| FP2005.1.2 | CGTGGCTCCCTCTACCAATTCTCC |
| FP2005.2.1 | AGCCATCCAGTAAGGGTTCCAAGC |
| FP2005.2.1 | AGCCATCCAGTAAGGGTTCCAAGC |
| FP2022.1.2 | GACCACAGCCTCCATTCACCATTC |
| FP2022.1.2 | GACCACAGCCTCCATTCACCATTC |
| FP2022.2.1 | CTCTCCGAGGCTTTGGGCTACAGT |
| FP2033.1.2 | GGGACCAAGAACCACAGACCTCCT |
| FP2033.1.2 | GGGACCAAGAACCACAGACCTCCT |
| FP2033.1.2 | GGGACCAAGAACCACAGACCTCCT |
| FP2033.2.1 | GCTGAGGGGAGAAACGCGAAATTA |
| FP2033.2.1 | GCTGAGGGGAGAAACGCGAAATTA |
| FP2033.2.1 | GCTGAGGGGAGAAACGCGAAATTA |
| FP2033.3.1 | AGAAACTGGGCGGTCTGAGTCTCC |
| FP2033.3.1 | AGAAACTGGGCGGTCTGAGTCTCC |
| FP2033.3.1 | AGAAACTGGGCGGTCTGAGTCTCC |
| FP2034.1.1 | TGTGAGCCTCGCCCTGCTAAATAA |
| FP2034.1.1 | TGTGAGCCTCGCCCTGCTAAATAA |
| FP2034.1.1 | TGTGAGCCTCGCCCTGCTAAATAA |
| FP2034.1.1 | TGTGAGCCTCGCCCTGCTAAATAA |
| FP2034.2.2 | CGCAGATGATGATTGTGGACCTGT |
| FP2034.2.2 | CGCAGATGATGATTGTGGACCTGT |
| FP2034.2.2 | CGCAGATGATGATTGTGGACCTGT |
| FP2034.2.2 | CGCAGATGATGATTGTGGACCTGT |
| FP2034.3.1 | TACTTGACACCCCAAGTCCAGTGC |
| FP2034.3.1 | TACTTGACACCCCAAGTCCAGTGC |
| FP2034.3.1 | TACTTGACACCCCAAGTCCAGTGC |
| FP2034.3.1 | TACTTGACACCCCAAGTCCAGTGC |
| FP2034.4.2 | GATTCCGCACTGGCAGAGAACCT |
| FP2034.4.2 | GATTCCGCACTGGCAGAGAACCT |
| FP2034.4.2 | GATTCCGCACTGGCAGAGAACCT |
| FP2034.4.2 | GATTCCGCACTGGCAGAGAACCT |
| FP2221.1.1 | GGAACAAGAAGATGGTGCGACGAC |
| FP2221.1.1 | GGAACAAGAAGATGGTGCGACGAC |
| FP2221.2.1 | AGAATCTCTTCAAGGGCGGAGCAC |

| | |
|------------|---------------------------|
| FP2266.1.2 | TCAAGCAATGGATGTGGATTTACCC |
| FP2266.1.2 | TCAAGCAATGGATGTGGATTTACCC |
| FP2266.2.1 | CAGACGTGAGCTGTAGCCGGACTT |
| FP2266.2.1 | CAGACGTGAGCTGTAGCCGGACTT |
| FP2356.1.1 | CGTGCTGGGAAATGTAGGCGATTA |
| FP2356.1.1 | CGTGCTGGGAAATGTAGGCGATTA |
| FP2356.2.1 | TGCTCATTCTGATCGGATGTGTCC |
| FP2423.1.2 | AAACTCAGAAGTGGGCCCCAAGAA |
| FP2423.1.2 | AAACTCAGAAGTGGGCCCCAAGAA |
| FP2423.1.2 | AAACTCAGAAGTGGGCCCCAAGAA |
| FP2423.2.3 | TGGAACCAGAACAGCAAAGGCCAAA |
| FP2423.2.3 | TGGAACCAGAACAGCAAAGGCCAAA |
| FP2423.2.3 | TGGAACCAGAACAGCAAAGGCCAAA |
| FP2423.3.2 | CTGCGTGCAAAAAGGAGAGTGACA |
| FP2423.3.2 | CTGCGTGCAAAAAGGAGAGTGACA |
| FP2486.1.2 | CAGCCTGTGAATGAGGTGGACCAT |
| FP2519.1.1 | CCGGAGGGTAGGGGGTAATCTCAT |
| FP2519.1.1 | CCGGAGGGTAGGGGGTAATCTCAT |
| FP2519.2.1 | CTTGCAGATCAGAACGGACCCTGT |
| FP2538.1.3 | CATGACATGGTACCTGCCTCTGGA |
| FP2538.1.3 | CATGACATGGTACCTGCCTCTGGA |
| FP2538.1.3 | CATGACATGGTACCTGCCTCTGGA |
| FP2538.2.3 | TGCTGAAAGAATGCACCCTGACAA |
| FP2538.2.3 | TGCTGAAAGAATGCACCCTGACAA |
| FP2538.2.3 | TGCTGAAAGAATGCACCCTGACAA |
| FP2538.3.2 | CATGTCACCACCCAAATGCTTGTC |
| FP2538.3.2 | CATGTCACCACCCAAATGCTTGTC |
| FP2538.3.2 | CATGTCACCACCCAAATGCTTGTC |
| FP2616.1.1 | GCTGACTGCTAACCACCTTCACCA |
| FP2616.1.1 | GCTGACTGCTAACCACCTTCACCA |
| FP2616.2.2 | TGAGGCATCATTTTAGGCCACAGG |
| FP2660.1.2 | GGAGTCCCTGGGGTTAAGAGGACA |
| FP2660.1.2 | GGAGTCCCTGGGGTTAAGAGGACA |
| FP2660.1.2 | GGAGTCCCTGGGGTTAAGAGGACA |
| FP2660.2.4 | TGGATAAAGCTCCGATTCCTGCTG |
| FP2660.2.4 | TGGATAAAGCTCCGATTCCTGCTG |
| FP2660.2.4 | TGGATAAAGCTCCGATTCCTGCTG |
| FP2660.3.1 | ATATCACAAAGCGTGCAGGCCAAG |
| FP2660.3.1 | ATATCACAAAGCGTGCAGGCCAAG |
| FP2808.1.4 | TTAAATTCGGGGCCGGTACACTTG |
| FP2808.1.4 | TTAAATTCGGGGCCGGTACACTTG |
| FP2808.2.1 | ACCACTTGACATTGAGGGGAAGA |
| FP2810.1.1 | CCATGGTGATTGCCCCTAGAAACA |

| | |
|------------|---------------------------|
| FP2810.1.1 | CCATGGTGATTGCCCTAGAAACA |
| FP2810.2.1 | TGACTGTCAGTGGAACAGCCAACC |
| FP2847.1.2 | CAGAAGCCACAGGATCCCAGATTG |
| FP3471.1.2 | TCTGGTGCTGATGAGATGGCTCTG |
| FP3471.1.2 | TCTGGTGCTGATGAGATGGCTCTG |
| FP3471.1.2 | TCTGGTGCTGATGAGATGGCTCTG |
| FP3471.2.3 | ACTGTGCTGGCTGGAAACCACTTC |
| FP3471.2.3 | ACTGTGCTGGCTGGAAACCACTTC |
| FP3471.2.3 | ACTGTGCTGGCTGGAAACCACTTC |
| FP3471.3.1 | TGCTTTGGGTAATGTCCGTTCTGG |
| FP3471.3.1 | TGCTTTGGGTAATGTCCGTTCTGG |
| FP3643.1.2 | TCGGCTTCTCACCGTGTTTGA |
| FP3643.1.2 | TCGGCTTCTCACCGTGTTTGA |
| FP3643.2.3 | GAGCACACATCCTCCACAGGACAA |
| FP3643.2.3 | GAGCACACATCCTCCACAGGACAA |
| FP4020.1.1 | ATCCAGAGGACCGTGCAACAAAAA |
| FP4185.1.2 | GACAAGCCGGACATAGGGGAAATC |
| FP4400.1.1 | TCAGCCTTTGCAGCGGAGAAAGTA |
| FP4400.1.1 | TCAGCCTTTGCAGCGGAGAAAGTA |
| FP4400.2.3 | CACTGCGGAGACCACTTCTTGTCC |
| FP4470.1.3 | GGAGAGTGGAAAGGGCCTTCATGT |
| FP4470.1.3 | GGAGAGTGGAAAGGGCCTTCATGT |
| FP4470.2.2 | GCCCTGATTTGAGCCCTTCAGTCT |
| FP4845.1.3 | GCCATGTGCTTACGCAGACAGGTT |
| FP4845.1.3 | GCCATGTGCTTACGCAGACAGGTT |
| FP4845.2.1 | CACAGGCCAACTTCTGCTTTACCG |
| RP81.2.2 | GGACTGTGCCTCCTTCTGGAACAA |
| RP81.3.3 | ACTGAGGGCCTTTTTTGATGCCAAC |
| RP81.3.3 | ACTGAGGGCCTTTTTTGATGCCAAC |
| RP330.3.4 | GGCTGTCCTGCTCTACCCGGATTA |
| RP330.4.1 | TTCATGGTTTTCTGGCGATCTGT |
| RP330.3.4 | GGCTGTCCTGCTCTACCCGGATTA |
| RP330.4.1 | TTCATGGTTTTCTGGCGATCTGT |
| RP400.3.2 | AGGCCCTTCCCTTGAGACTCTGTG |
| RP400.4.2 | CAGTCCTAGCTGAGGATGGGGACA |
| RP400.3.2 | AGGCCCTTCCCTTGAGACTCTGTG |
| RP400.4.2 | CAGTCCTAGCTGAGGATGGGGACA |
| RP455.2.1 | TGTCGGTCTGTCTCTCTGGGCTTC |
| RP455.3.1 | AGGTGGGCTTTTGTCAAGGATGGT |
| RP455.3.1 | AGGTGGGCTTTTGTCAAGGATGGT |
| RP467.3.5 | TTCTGCAATTTCTGGGCAACTAGGG |
| RP467.4.2 | ATGGCACCAGGTCAATAAGGTTGC |
| RP467.5.1 | TCCTGGAAATGTGCAGATGGATTG |

| | |
|------------|---------------------------|
| RP467.3.5 | TTCTGCAATTTCTGGGCAACTAGGG |
| RP467.4.2 | ATGGCACCAGGTCAATAAGGTTGC |
| RP467.5.1 | TCCTGGAAATGTGCAGATGGATTG |
| RP467.4.2 | ATGGCACCAGGTCAATAAGGTTGC |
| RP467.5.1 | TCCTGGAAATGTGCAGATGGATTG |
| RP486.3.5 | CAAGACATGGGGGACAAAGAACGA |
| RP486.4.1 | CAGCCTCAGCATTTTCCTGGTCTGT |
| RP486.5.1 | CCCTCCCTGAGCTGTTAGGTCCTG |
| RP486.3.5 | CAAGACATGGGGGACAAAGAACGA |
| RP486.4.1 | CAGCCTCAGCATTTTCCTGGTCTGT |
| RP486.5.1 | CCCTCCCTGAGCTGTTAGGTCCTG |
| RP486.4.1 | CAGCCTCAGCATTTTCCTGGTCTGT |
| RP486.5.1 | CCCTCCCTGAGCTGTTAGGTCCTG |
| RP724.3.1 | CGGGCCATGTCTTACTGTTCGATGT |
| RP724.4.1 | ACCCAGCTTCGTTCTCCTATGCTG |
| RP724.3.1 | CGGGCCATGTCTTACTGTTCGATGT |
| RP724.4.1 | ACCCAGCTTCGTTCTCCTATGCTG |
| RP757.3.1 | TGAGCTAGAAGGGACCCATGGACA |
| RP757.4.4 | CTGGCTTCGCCTTCAGCTTTGTAA |
| RP757.3.1 | TGAGCTAGAAGGGACCCATGGACA |
| RP757.4.4 | CTGGCTTCGCCTTCAGCTTTGTAA |
| RP869.2.1 | TGCGTGTCCCGAGAATAGAAAGGA |
| RP869.3.1 | AAGGCCTAGGCAGGAAGGCAATTT |
| RP869.3.1 | AAGGCCTAGGCAGGAAGGCAATTT |
| RP947.3.1 | CTGTCAGCTCGCAGTTCAAGGTCA |
| RP947.4.2 | CTGCTTGCCCACTCTATGGTCGTT |
| RP947.3.1 | CTGTCAGCTCGCAGTTCAAGGTCA |
| RP947.4.2 | CTGCTTGCCCACTCTATGGTCGTT |
| RP978.2.1 | AGCATCTCCATTTTCCCAGGCTGT |
| RP978.3.3 | TGTCAGTGCACGTTTACAGCAGCA |
| RP978.3.3 | TGTCAGTGCACGTTTACAGCAGCA |
| RP1004.4.1 | GGCTTTCCAGATCCAGTGTGAGGA |
| RP1004.5.2 | TGTAAGCCCCTGAGTTAGGCAGCA |
| RP1004.6.3 | GTCAAGACTCCCTCCGCCTTAGGA |
| RP1004.4.1 | GGCTTTCCAGATCCAGTGTGAGGA |
| RP1004.5.2 | TGTAAGCCCCTGAGTTAGGCAGCA |
| RP1004.6.3 | GTCAAGACTCCCTCCGCCTTAGGA |
| RP1004.4.1 | GGCTTTCCAGATCCAGTGTGAGGA |
| RP1004.5.2 | TGTAAGCCCCTGAGTTAGGCAGCA |
| RP1004.6.3 | GTCAAGACTCCCTCCGCCTTAGGA |
| RP1113.3.2 | GCTCAGAGCCCGTTCCTGGTTTAG |
| RP1113.4.1 | TGTCCGGAAAGGTTTTCTCCTGGT |
| RP1113.5.1 | AAGACATCACCAGGCAGCATCTCA |

| | |
|------------|----------------------------|
| RP1113.3.2 | GCTCAGAGCCCGTTCCTGGTTTAG |
| RP1113.4.1 | TGTCCGGAAAGGTTTTCTCCTGGT |
| RP1113.5.1 | AAGACATCACCAGGCAGCATCTCA |
| RP1113.4.1 | TGTCCGGAAAGGTTTTCTCCTGGT |
| RP1113.5.1 | AAGACATCACCAGGCAGCATCTCA |
| RP1131.2.3 | CGAGAATCTGCAGCTGTGTCAGGA |
| RP1131.3.1 | TTTTTCACCGCTCTGGAAGATGGA |
| RP1131.3.1 | TTTTTCACCGCTCTGGAAGATGGA |
| RP1153.3.2 | GCCCGACATTAATCCGCAGTCTTT |
| RP1153.4.1 | AAACCTTAGGGCCAAGCGGAGACT |
| RP1153.3.2 | GCCCGACATTAATCCGCAGTCTTT |
| RP1153.4.1 | AAACCTTAGGGCCAAGCGGAGACT |
| RP1205.2.4 | CTTGGTCCAGCCATGGCAAACCTTA |
| RP1259.2.1 | GCACTCACATTCTCTCCACGTCA |
| RP1259.3.1 | GCAATTCAAAGGAATGACCCAGCTC |
| RP1259.3.1 | GCAATTCAAAGGAATGACCCAGCTC |
| RP1450.2.1 | TGAAGGGCAGCATCTCGTGAGAAA |
| RP1450.3.1 | CTGCCGTTTAAACTGTGCATCGTG |
| RP1450.3.1 | CTGCCGTTTAAACTGTGCATCGTG |
| RP1520.3.2 | ACTCTTGGTGGGAGCAGGTGGTTT |
| RP1520.4.2 | CTGGACACCCAGTGCATGAGGAT |
| RP1520.3.2 | ACTCTTGGTGGGAGCAGGTGGTTT |
| RP1520.4.2 | CTGGACACCCAGTGCATGAGGAT |
| RP1541.2.1 | CACAGAGGGAATCCACTGGTGGTC |
| RP1541.3.1 | TGTTGTGGCCACTGGCTTGTTAGA |
| RP1541.3.1 | TGTTGTGGCCACTGGCTTGTTAGA |
| RP1581.3.1 | GAGGAAGATGCTATCAGAAGGGTTGA |
| RP1581.4.1 | GGAGGTGCTGTTGAGGTCGTCAGT |
| RP1581.5.3 | GTCACCAGTCCTATGTCCCCACGA |
| RP1581.3.1 | GAGGAAGATGCTATCAGAAGGGTTGA |
| RP1581.4.1 | GGAGGTGCTGTTGAGGTCGTCAGT |
| RP1581.5.3 | GTCACCAGTCCTATGTCCCCACGA |
| RP1581.4.1 | GGAGGTGCTGTTGAGGTCGTCAGT |
| RP1581.5.3 | GTCACCAGTCCTATGTCCCCACGA |
| RP1590.2.4 | AGAAGCTGTGGTGATCGGGCTTTT |
| RP1590.3.1 | CTCACTGCACAAACAGCGAGTGGT |
| RP1590.3.1 | CTCACTGCACAAACAGCGAGTGGT |
| RP1647.4.1 | AAGCCAAAGACACCAGGGTGTTGA |
| RP1647.5.2 | CTGTGTGATCCAGGGTGGGTGTCT |
| RP1647.6.2 | GAATTCCCCGTCTTGACAATGCAC |
| RP1647.7.1 | AGCACATTAGCAGGTCAACCAGGA |
| RP1647.4.1 | AAGCCAAAGACACCAGGGTGTTGA |
| RP1647.5.2 | CTGTGTGATCCAGGGTGGGTGTCT |

| | |
|------------|---------------------------|
| RP1647.6.2 | GAATTCCCCGTCTTGACAATGCAC |
| RP1647.7.1 | AGCACATTAGCAGGTCAACCAGGA |
| RP1647.4.1 | AAGCCAAAGACACCAGGGTGTGTA |
| RP1647.5.2 | CTGTGTGATCCAGGGTGGGTGTCT |
| RP1647.6.2 | GAATTCCCCGTCTTGACAATGCAC |
| RP1647.7.1 | AGCACATTAGCAGGTCAACCAGGA |
| RP1647.5.2 | CTGTGTGATCCAGGGTGGGTGTCT |
| RP1647.6.2 | GAATTCCCCGTCTTGACAATGCAC |
| RP1647.7.1 | AGCACATTAGCAGGTCAACCAGGA |
| RP1688.2.1 | GGACATGTGTTTCCTCCTCCGTTG |
| RP1688.3.2 | GGGTTGGGTCTGGCGTCTAGTTTC |
| RP1688.3.2 | GGGTTGGGTCTGGCGTCTAGTTTC |
| RP1753.3.2 | CTGGACAACAGCACAAAGGCACACT |
| RP1753.4.1 | CACGTTTGTGTGCCATTGGAGAAG |
| RP1753.5.3 | CACGGGGTGAAGAGGAGAGTGTGT |
| RP1753.3.2 | CTGGACAACAGCACAAAGGCACACT |
| RP1753.4.1 | CACGTTTGTGTGCCATTGGAGAAG |
| RP1753.5.3 | CACGGGGTGAAGAGGAGAGTGTGT |
| RP1753.4.1 | CACGTTTGTGTGCCATTGGAGAAG |
| RP1753.5.3 | CACGGGGTGAAGAGGAGAGTGTGT |
| RP1777.2.1 | TCCACACCCACAATACTCGGCTTT |
| RP1777.3.1 | TGATGTCTGGAGGAGTGCCATCAG |
| RP1777.3.1 | TGATGTCTGGAGGAGTGCCATCAG |
| RP1928.3.3 | ACTGCGCTTCTCGAGTTTCACACC |
| RP1928.4.4 | GCTTGAGCTTGCACCAAGTTGCTC |
| RP1928.3.3 | ACTGCGCTTCTCGAGTTTCACACC |
| RP1928.4.4 | GCTTGAGCTTGCACCAAGTTGCTC |
| RP2005.3.2 | TGTGGGCAGTAGGAAAGGCAGAAC |
| RP2005.4.2 | CCACAGAGGGCTCACGGTAATGAA |
| RP2005.3.2 | TGTGGGCAGTAGGAAAGGCAGAAC |
| RP2005.4.2 | CCACAGAGGGCTCACGGTAATGAA |
| RP2022.2.1 | ACTGTAGCCCAAAGCCTCGGAGAG |
| RP2022.3.2 | TGTCCGGTTTGATCATTGCTGTGT |
| RP2022.3.2 | TGTCCGGTTTGATCATTGCTGTGT |
| RP2033.4.1 | TATTCAGGTGGAGTGCAACGTGGA |
| RP2033.5.4 | GACCGAGAGACGCTTGGTTGAAGA |
| RP2033.6.3 | GAGTCCGGAGATGGGAACAACACA |
| RP2033.4.1 | TATTCAGGTGGAGTGCAACGTGGA |
| RP2033.5.4 | GACCGAGAGACGCTTGGTTGAAGA |
| RP2033.6.3 | GAGTCCGGAGATGGGAACAACACA |
| RP2033.4.1 | TATTCAGGTGGAGTGCAACGTGGA |
| RP2033.5.4 | GACCGAGAGACGCTTGGTTGAAGA |
| RP2033.6.3 | GAGTCCGGAGATGGGAACAACACA |

| | |
|------------|---------------------------|
| RP2034.5.2 | GGCCAGGTTCCCTCTCTGTGCTTCT |
| RP2034.6.4 | GCAGGGATTTGGAAGGATGTCTGA |
| RP2034.7.2 | GGTGACCTGAAGATCAGGCAGGAG |
| RP2034.8.1 | GGGGAAATACAGAGCCCCATCTGA |
| RP2034.5.2 | GGCCAGGTTCCCTCTCTGTGCTTCT |
| RP2034.6.4 | GCAGGGATTTGGAAGGATGTCTGA |
| RP2034.7.2 | GGTGACCTGAAGATCAGGCAGGAG |
| RP2034.8.1 | GGGGAAATACAGAGCCCCATCTGA |
| RP2034.5.2 | GGCCAGGTTCCCTCTCTGTGCTTCT |
| RP2034.6.4 | GCAGGGATTTGGAAGGATGTCTGA |
| RP2034.7.2 | GGTGACCTGAAGATCAGGCAGGAG |
| RP2034.8.1 | GGGGAAATACAGAGCCCCATCTGA |
| RP2034.5.2 | GGCCAGGTTCCCTCTCTGTGCTTCT |
| RP2034.6.4 | GCAGGGATTTGGAAGGATGTCTGA |
| RP2034.7.2 | GGTGACCTGAAGATCAGGCAGGAG |
| RP2034.8.1 | GGGGAAATACAGAGCCCCATCTGA |
| RP2221.2.2 | CCGCCCTTGAAGAGATTCTGTGTG |
| RP2221.3.2 | GCGGAGGGAGGGAGCTTTATCTTT |
| RP2221.3.2 | GCGGAGGGAGGGAGCTTTATCTTT |
| RP2266.3.2 | TCGCAGTCTGGGGGAATAAACTCA |
| RP2266.4.1 | TGTGTCCAAAAGTCCAGGTGTCCA |
| RP2266.3.2 | TCGCAGTCTGGGGGAATAAACTCA |
| RP2266.4.1 | TGTGTCCAAAAGTCCAGGTGTCCA |
| RP2356.2.2 | CCGATCAGAATGAGCAGTCCATGA |
| RP2356.3.2 | GCATCAAACATTACGGATGTCCA |
| RP2356.3.2 | GCATCAAACATTACGGATGTCCA |
| RP2423.3.1 | CCAAACATTCCAAGCCAAGATCCA |
| RP2423.4.1 | AGTTCCTGGCTCCGTGCCTTATGT |
| RP2423.5.2 | AAGTGTGTCTGGCTAGGGGATCCTG |
| RP2423.3.1 | CCAAACATTCCAAGCCAAGATCCA |
| RP2423.4.1 | AGTTCCTGGCTCCGTGCCTTATGT |
| RP2423.5.2 | AAGTGTGTCTGGCTAGGGGATCCTG |
| RP2423.4.1 | AGTTCCTGGCTCCGTGCCTTATGT |
| RP2423.5.2 | AAGTGTGTCTGGCTAGGGGATCCTG |
| RP2486.2.1 | AGAGCTTACTCCACCTGCCGTCTT |
| RP2519.2.2 | TTCGGCCTCCGAAGTTCTCCCTAT |
| RP2519.3.1 | TGCTTGGTCAGTCAGCCTCCCTTA |
| RP2519.3.1 | TGCTTGGTCAGTCAGCCTCCCTTA |
| RP2538.4.1 | ATTGTTCCGAGCCATGCAGATGAG |
| RP2538.5.4 | CAGGCTCACGGACTGCATTGTTTT |
| RP2538.6.4 | TCCCACGCAGTGTGTCTAGTGAA |
| RP2538.4.1 | ATTGTTCCGAGCCATGCAGATGAG |
| RP2538.5.4 | CAGGCTCACGGACTGCATTGTTTT |

| | |
|------------|---------------------------|
| RP2538.6.4 | TCCCACGCAGTGTGTCCTAGTGAA |
| RP2538.4.1 | ATTGTTCCGAGCCATGCAGATGAG |
| RP2538.5.4 | CAGGCTCACGGACTGCATTGTTTT |
| RP2538.6.4 | TCCCACGCAGTGTGTCCTAGTGAA |
| RP2616.2.4 | GGTTCCTTTGGCCGATGTCTTCAT |
| RP2616.3.1 | GTGCAGCGATAAATGAGGGACGAC |
| RP2616.3.1 | GTGCAGCGATAAATGAGGGACGAC |
| RP2660.3.1 | GAAGTTCATTGGCCCCACACCTGAG |
| RP2660.4.4 | GCTCCATGAGTGCTCCATGATGTG |
| RP2660.5.3 | CACAAAGGGTGTCCAAGGTTCCAG |
| RP2660.3.1 | GAAGTTCATTGGCCCCACACCTGAG |
| RP2660.4.4 | GCTCCATGAGTGCTCCATGATGTG |
| RP2660.5.3 | CACAAAGGGTGTCCAAGGTTCCAG |
| RP2660.4.4 | GCTCCATGAGTGCTCCATGATGTG |
| RP2660.5.3 | CACAAAGGGTGTCCAAGGTTCCAG |
| RP2808.2.1 | TCTTCCCCTCAATGTGCAAGTGGT |
| RP2808.3.1 | AGAAACCCTGGCAAGAGGACAAGG |
| RP2808.3.1 | AGAAACCCTGGCAAGAGGACAAGG |
| RP2810.2.2 | ATTGGGTTGGCTGTTCCACTGACA |
| RP2810.3.4 | TGCTTTGGGTGTGAGGTTGGACTT |
| RP2810.3.4 | TGCTTTGGGTGTGAGGTTGGACTT |
| RP2847.2.5 | GAAAGGCTCATGGGCATTGAACAC |
| RP3471.3.1 | CCAGAACGGACATTACCCAAAGCA |
| RP3471.4.2 | GCCAGAATACAGGTCAGCCTGTGC |
| RP3471.5.3 | ATGATGATGCAGTCTGGACGCAAA |
| RP3471.3.1 | CCAGAACGGACATTACCCAAAGCA |
| RP3471.4.2 | GCCAGAATACAGGTCAGCCTGTGC |
| RP3471.5.3 | ATGATGATGCAGTCTGGACGCAAA |
| RP3471.4.2 | GCCAGAATACAGGTCAGCCTGTGC |
| RP3471.5.3 | ATGATGATGCAGTCTGGACGCAAA |
| RP3643.3.1 | CAGGTCAGGTCAGAACGGAGGCTA |
| RP3643.4.1 | GTATGCCAGGCGCTATACGCAAGA |
| RP3643.3.1 | CAGGTCAGGTCAGAACGGAGGCTA |
| RP3643.4.1 | GTATGCCAGGCGCTATACGCAAGA |
| RP4020.2.4 | TCCGGCTGATGATGAACTGATTGA |
| RP4185.2.2 | CTTGTGGCTCGGGTCCATCTTACA |
| RP4400.2.2 | GGACAAGAAGTGGTCTCCGCAGTG |
| RP4400.3.3 | CGACATGGCTCTGGGCATATGTT |
| RP4400.3.3 | CGACATGGCTCTGGGCATATGTT |
| RP4470.2.3 | GAGCCACAGACTGAAGGGCTCAAA |
| RP4470.3.1 | CTTCCTTGGATGGAGATCGGGTGT |
| RP4470.3.1 | CTTCCTTGGATGGAGATCGGGTGT |
| RP4845.2.1 | CGGTAAAGCAGAAGTTGGCCTGTG |

| | |
|------------|--------------------------|
| RP4845.3.2 | AGCTCAAGCATGGCGGTTATGATG |
| RP4845.3.2 | AGCTCAAGCATGGCGGTTATGATG |

Primers used to amplify novel 3' exons on RefSeq genes

| | |
|----------------------------|----------------------------|
| RP1785.1.1 | GAGGCACGTCCTAATCCACACTGG |
| RP657.1.1 | AGATGGAGGGTGTCCCGACTTCTC |
| RP1576.1.4 | GTGAGGCTCTTTTGGGGACATCAC |
| RP2518.1.1 | ACACATCGGACACCTTGTGCCTTT |
| RP3522.1.1 | AGAGCGGTAATGCAGCTGAACTCG |
| RP4906.1.3 | AGTGAGGCACGCAGAAATCCAGTT |
| RP5032.1.1 | GGGTCGAGGATTTTTAGGGATGGA |
| RP688.1.3 | ACATCCTAAGCGCTGGTTCCCCTA |
| RP1778.1.1 | ACAGAACCCCGTGGAGTACAAGCA |
| RP1600.1.3 | TGTTCTTCCGTAGGGCACCTCAGT |
| forward | forward_seq |
| FPENSMUSG00000001281.15.1 | GTGCGATCACAACCACTGTCAACC |
| FPENSMUSG000000020839.13.1 | ATGCCAGGGAGCCAATAAAGATGC |
| FPENSMUSG000000022148.21.3 | TCTGCTCTGTTTCACTTCCACTGTGC |
| FPENSMUSG000000025035.6.4 | CCGGGAGCTAGAGTCAGCCCTAGA |
| FPENSMUSG000000028982.11.2 | TACCTGCTGGGAGAGCGTGCTTAG |
| FPENSMUSG000000032175.23.1 | TACTCTGCCTGGAAACCCACCT |
| FPENSMUSG000000032491.5.3 | AGATCCGGCCACTTCATGTTTCCTT |
| FPENSMUSG000000033983.7.1 | ATAGAGCATCTCGCCCATTCACA |
| FPENSMUSG000000037525.2.3 | GGATGGAGTTAGCGTGCTGTTTCG |
| FPENSMUSG000000038725.78.1 | GTTCGTGTGGATTCACGACCCTTC |

Primers used to amplify novel 5' exons on Ref Seq genes

| | |
|---------------------------|--------------------------|
| FP3061.1.1 | GCCTGACCCACAGACCAACTGACT |
| FP1653.1.2 | GCAGGTGAACAATCGTTGTGATCG |
| FP2779.1.2 | GGGGAATGGAAGCAGTCCTAGGTG |
| FP864.1.1 | CGGGGCTTACCTGAAGCTATGGAG |
| FP4957.1.3 | AGGTGACAGTGGAACCTGCAGACC |
| FP4297.1.1 | CCTTCAGCCCAAATGCTTGTCATC |
| FP497.1.3 | CCCCTGAATTCCAAGTGTGGTCTC |
| FP1577.1.2 | AGGACCAGGGAAACGAACCTACCC |
| FP4957.1.3 | AGGTGACAGTGGAACCTGCAGACC |
| FP3572.1.1 | GCAGTCTCCTTCCATCCATCGTTC |
| RPENSMUSG000000053819.1.1 | TTCCGAGCTCCTCAAAGAGCTGAT |
| RPENSMUSG000000064210.1.2 | GTCTCCATCCTCATCGTCGTCCTC |
| RPENSMUSG000000039483.1.3 | TTCCGCACCGGAAGTTATCCTACC |
| RPENSMUSG000000032782.1.1 | GAACGGTTAAAAGCGGATGTGCAA |
| RPENSMUSG000000032733.1.1 | AATCATAGAGGGCTCGGCCTTTC |
| RPENSMUSG000000030965.1.1 | GAAGGTGTAGCCCGAAATGGAAGC |

| | |
|--------------------------|---------------------------|
| RPENSMUSG00000055053.1.1 | GATGAACGGGTGGAACATCATCCTG |
| RPENSMUSG00000054263.1.1 | CAGGAGGGTCAGAGCTGACAGGAG |
| RPENSMUSG00000032733.1.1 | AATCATAGAGGGCTCGGCCTTCA |
| RPENSMUSG00000029195.1.5 | TGCCAAGCAACAAGGTAAGGGTTG |

Primers used to amplify internal exons on RefSeq genes

| | |
|-------------|----------------------------|
| FP195.4.5 | CCAACAGCCTAATGCTGAAGCACA |
| FP234.3.1 | GGCCCTGAACAGAAAACCTGGAAG |
| FP355.4.1 | AGTCCCTGGGCATTCCTCAGAAAA |
| FP6974.3.1 | TGTGTGGAGCACCATACCTACCACA |
| FP4224.1.1 | CCCGGAACCATGAACCCTAACTGT |
| FP4533.1.1 | TGCAAAAATACCAGTCCCCAGTGC |
| FP4591.3.3 | TCTCAGCCATTTTGCACAGACCAG |
| FP10649.3.3 | TCGTAGCCCTACTCTGTGCCCTTG |
| FP14909.2.3 | CCTTGAGACCACGTCTCTGCTTCC |
| FP15423.2.2 | TCCAGGAAACAGATCCTCGACTGG |
| FP15030.2.1 | AATGTTCACTCACACCGGGCAGTT |
| FP18233.4.1 | CTTCGGTCCCTTTAGCCGTTCTTG |
| FP19324.1.4 | TTCGGAGGTCTGGACAGACTAGCA |
| FP20705.2.1 | TGGCAGCAAAAATTCCCTTCTGA |
| FP23834.3.1 | TAGGCACCATCTTGTAGCCCTGGA |
| FP29245.3.2 | GGTTCATCCCAAACTGATGAGCA |
| FP29854.3.2 | AGAAGCCTCCTTCACTCCCCAGGT |
| FP9538.2.3 | GCTGCGACTTGCAGTCGATGGTAT |
| FP31215.3.1 | CTGTGAAGTTCCATGCCAGGACAG |
| FP23698.1.1 | ATGGGAAAAAGCAGTGGGATTTGG |
| RP195.5.2 | TTAGCATCAGCGACAGCCAGAGGT |
| RP234.4.1 | TCTGCTTCCCGTCTTCATAGTGGA |
| RP355.5.3 | CCACTCTTCCTTCATGGGTGCAAG |
| RP6974.4.1 | TTGGAGATCATGGAAGTGGCTCGT |
| RP4224.2.2 | AACTTTGCCACACCCAGGTCTCT |
| RP4533.2.1 | CTTGCTCTAACACAGCAGCAGCA |
| RP4591.4.1 | CTATGGGCCTCGATGCATGATCTC |
| RP10649.4.1 | ACCTGATTCGCTGGCGTAGAGATG |
| RP14909.3.1 | ACCTGGGGAGGAACACACTTTCCA |
| RP15423.3.2 | GATACCATGCAGTGCAAAGCACCT |
| RP15030.3.2 | TAAGCTGTGTGCAGTCTGAAGCAA |
| RP18233.5.1 | GAGTGTACCCTGCCGGCTTCTTCT |
| RP19324.2.4 | TTAGAAGGGCTTTGGGGGATGGTT |
| RP20705.3.1 | CCTAGGAAGCGAGGGGTCTGGTTC |
| RP23834.4.1 | TCGATCTTGCTGGACCACTTCTCC |
| RP29245.4.2 | GCATGTTTCCTCTTCCGTTTCGAAAA |
| RP29854.4.3 | TCTGTCTGTCAGCCATCAACAGCA |

| | |
|-------------|--------------------------|
| RP9538.3.1 | CGCTGAGAGACACCATCACAAAGC |
| RP31215.4.5 | GTATGCTGTTCTCCTGGGCCATGT |
| RP23698.2.1 | AGGCACAGCTGTAGGTTGGTTTCG |

Primers on hypertrapped genes

| | |
|-----------|----------------------------|
| FPE577769 | GCCATACACCATGGATGCGTTC |
| FPE141136 | GCACACCTTACGGACACGGAGA |
| FPE214933 | GGCCATCCATAACCGAGGGAAA |
| FPE110463 | ACTAGCGCCACCGCCCTTTCT |
| FPE277766 | TTGGACAGGCGCATGGTTAAGG |
| FPE352331 | CGGCAAAGCCGAGAAGGAGAAC |
| FPE250850 | CCCTACAGTGGCTGTGGGAAAGTC |
| FPE106542 | GCTTCCGACATGATGGTTCTCCTG |
| FPE295390 | GATGACCCGCAGAGTGGAGAGC |
| FPE214915 | TGGAATCCGCGAAGATCAGAGC |
| RPE577768 | AGCAAATCCGGGGTAGCCTCTG |
| RPE511104 | GCTGGTTGTGAATTACTTCCTTGG |
| RPE392401 | CTGCCTGTGGTCCACTCGATCC |
| RPE396136 | GAAGATGGCCTGCCACTCAGGA |
| RPE127477 | CCAAGTTCTTCTCAGGTTCCCAAG |
| RPE582536 | GGGCTTGATGTCCAGAGGCAGA |
| RPE362649 | CTGGCCACGTGCAGGGAAAG |
| RPE264351 | CCAGCAGCCATCTTTCCTCCGTA |
| RPE295384 | GGTGTGTTGGGCTCATTAAGCAGTGA |
| RPE214916 | TGTCAACAATCTGAGAGCCCGAGA |

Primers on trapped RefSeq

| | |
|---------------|---------------------------|
| NP_808265.1 | CACCAGCACCATCAGCCCATTT |
| Cklfsf5 | ACCAGCGCTTCGATAGGCTCAA |
| Gcnt3 | CTCCACATCACTCACGGCGTTG |
| Cbr1 | AATACGGAGGCCTGGACGTGCT |
| Ly6g6c | TCCTGTTGCTCACCCGTGTCTGC |
| Zik1 | TGCAGAAATGGATATGGCCCTCA |
| Gprc5b | CCAGTGCACCGTTCAGAAGCAA |
| Egr1 | GGAGCCGAGCGAACAACCCTAT |
| Tceal1 | CCTTGATCGAGAAGGAAAGCAGAA |
| 2610319K07Rik | CGCCGTA CTTG CAGGAAAGCAG |
| RPE365971 | AGACCCATCACATCGGCAAGGA |
| RPE320895 | GCAGCAAAGGAGACCACCAGGA |
| RPE217909 | CATCAAGCCTTGCCCAGCAGAG |
| RPE253497 | TTGAATGTGGAAGGGGGTGTCTG |
| RPE141753 | ACAGCCCAGCACAGGGACTTTG |
| RPE198304 | TCTCCATTCTTCATGGGAGAAGCAA |

| | |
|-----------|-------------------------|
| RPE410397 | GTGTGAGATGGCGGAGCAGTTG |
| RPE433523 | TCGTCTCCACCATCGCCTTCTC |
| RPE389668 | CTCTTCATTTTCACTGCGCGTGT |
| RPE177331 | CGGGAATACCACTTGTTCTGAC |

Primers used to amplify fragments tested in zebrafish

| | |
|--------|-------------------------|
| 2894F | CAAATGACAGACGACACCTAAG |
| 2894R | TTCTCTTTGTGGTCCCTGCT |
| 2755F | GGTATCTGTGCGCCTTTTCT |
| 2755R | CAGATTTGAATTTGCAGCGA |
| 2756F | CGCAGATGAAATTGGACAGA |
| 2756R | GACAGGACATCAGGACAGGC |
| 1645F | GGGATGTGTTCTCCATGCTT |
| 1645R | CAATACAATGACGGAGAGGG |
| 1646F | CCGATTCTCCCATCAGTTCA |
| 1646R | CTGTAGTGGGGCAACAGGAG |
| 1652F | ACATTCAGAAGAGCCAGCGT |
| 1652R | GCAGCCATAGTTCCCAGTCT |
| 1653F | AGCCTAAACACACCACCTCG |
| 1653R | CGTGAGAAAATGGCTGACGTA |
| 1653IF | GTCCCGGTACACAACAAGGA |
| 1553IR | GACATTCTGGAACCCTCCAA |
| 333F | AACCACGAGTGAAACTCGGA |
| 333R | CATTCAGCCTGGCTCTCTGT |
| 1194F | TGACACAACGGGAAACTACA |
| 1194R | GAACTGGGAAGTGTGCAAGG |
| 2598F | ATACCCCTGGGTAAAAGGC |
| 2598R | GCTGCTGAACAGCGGTAAGT |
| 2598IF | TGCACCTCTGATTGAGGAGTT |
| 2598IR | CCCCTGTCTTTATGAGATAACCA |
| 44F | CACGTGTTGTGCTTGTTTCC |
| 44R | TCCTTTACCTTCCAACCCTG |
| 45F | CCCTAGGAGGGGTCTCAGTAG |
| 45R | ATGCTTCCATCTGCTGGTCT |
| 691F | GCATACAGTCGCCAAACTC |
| 691R | ACGCTTAGGTATCAGCGGAG |
| 692F | GCTTGTTGACGGAGTGGTTC |
| 692R | CAGAACGCTGCTTTGTGAGA |
| 1050F | TAGCCTGATGGCCATTAACA |
| 1050R | GGAAACTTACATTCGCTCCC |
| 1051F | AGATAGCACAGGCCAGATGAA |
| 1051R | GGAGCGATAAAAAGATGAGCA |

| | |
|-------|-----------------------|
| 1052F | TGGGGAGAAACAATGTAGGC |
| 1052R | GAAACCCTCCTCCATGATCC |
| 3120F | AAGACTCTTTTGGTGGCCTCT |
| 3120R | TCAAGAGGGTGATGGTCTCTG |
| 3121F | ATCCAATTTGGCTGATCCTG |
| 3121R | TGTTGTCTGACGCTCAATGC |
| 3122F | GTGCGCTCATTACCAGCTCT |
| 3122R | GGAGTACAAGTGCAACTGCC |
| 1972F | TTAGGCACTGGGGACAGAGT |
| 1972R | TTGATGGTGGTTGCATTTTG |
| 1973F | GTGAGGCAGGCTTGGTTAAG |
| 1973R | AAGTCTCCTGGAGCCATCTTC |
| 4939F | GGTGGTTTTCCCTGACTGAG |
| 4939R | TGGCAGTCTAGAGCCACCAT |
| 4940F | GGGCAATATCAGAGCGAGAC |
| 4940R | CCCTCAAATGTTCTGTTCTG |
| 2032F | TGGCTTTTGTCCATTCTGTG |
| 2032R | ATCCCTAACCCAACATTCTGT |
| 4049F | GTCACCCGCTGTTATGATTG |
| 4049R | GTTCTGCCCACGACTGATCT |

Primer used to amplify negative controls fragments

| | |
|----------|-----------------------|
| VC11216F | TCTGCTTACAGATGCTGGCT |
| VC3255F | CCCATAATGGACACCCTCTG |
| VC2797F | GGTGGTCGGCTGTAAAAAGA |
| VC198F | TGCTTAGTTTGTGTCTGGTGG |
| VC909F | GTGTGTCATCCTCATCCACG |
| VC410F | GGAAGCCTTTTTACCCCAGA |
| VC10157F | TTGGAGATCAGATCACAGGG |
| VC11271F | GCCGTTGCGTTTTATTTAGC |
| VC5990F | CAGTGTTGCAGGCAGAAGAA |
| VC268F | TCTTCTTTCCGTCGAACTGG |
| VC11767F | CCTTTCATACGTCGCTCGAT |
| VC5945F | GTCCGCGAGTGCAATAAATC |
| VC11216R | CTCGTAAAGGGTGTGGTGTG |
| VC3255R | GTTCTGGACGCATCAGGATT |
| VC2797R | CGGTGGTCCCTATCTGAGTC |
| VC198R | TCCTCCATTTTGTGGTCC |
| VC909R | CATTCCATGATGGTGCTCTG |
| VC410R | TCATATCCAAACCCGAGGAG |
| VC10157R | CGATGGATGAATCAGCAAGA |
| VC11271R | CAGCTGCTAAGGATCATGGG |
| VC5990R | CAAGGGAACACGGGGTATTA |

| | |
|----------|----------------------|
| VC268R | TCCGGATGATGGTGTACAG |
| VC11767R | CGTGTTCCTAAACACAACC |
| VC5945R | TGCAGAGGTCACAGAAATGC |

Primer used to amplify positive controls fragments

| | |
|----------|-----------------------|
| VF17653F | CTGTCGGTCAGACTCCAACA |
| VF17654F | AGTTCACCTGGTGTGCTGAA |
| VF17655F | CATCAGTGACACATGGCGTT |
| VF17656F | GGCCCCATAAAGGTTTCATTC |
| VF5490F | GTGTGGTGACTCAGCCATGT |
| VF5491F | GCCATTGATTCCCTCCAGAC |
| VF5492F | GTTTAGCTCCAGCCCTGATG |
| FF6026F | CCAATATTTCCCATCAGCCT |
| FF12058F | GTGTCGACTCGAGGTGAAGA |
| FF12057F | ACACGCTTCCTGGAGATGAC |
| FF28050F | GACGTCACCGTGGAATTGTC |
| VF17653R | CAGCAATGGAAGCAGTGAGA |
| VF17654R | AGCCATTTCTTTCGTTACGG |
| VF17655R | GGAAGGAACTGTGGGAATG |
| VF17656R | CCCTCCATAAACAGCTCACA |
| VF5490R | GGACAACCTGGGATTGTGGTT |
| VF5491R | GGAAGGGAATTTTGGCACTT |
| VF5492R | GAGCACTAACCTCGACAGGC |
| FF6026R | GCACGCGTCACAATGTCTTA |
| FF12058R | GCTTGAATCGAGGTCTCAGC |
| FF12057R | CTGAAGACAGACTCCGTCCC |
| FF28050R | AATTAGGCCGAAGGGGTAAG |